UNIVERSITÉ SAVOIE MONT BLANC

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ SAVOIE MONT BLANC

Spécialité : **SCIENCES ET TECHNIQUES DE L'INFORMATION ET DE LA COMMUNICATION, TRAITEMENT DE L'INFORMATION**

Arrêté ministériel : 25 Mai 2016

Présentée par

Antoine BRALET

Thèse dirigée par **Abdourrahmane M. ATTO**,
Et codirigée par **Emmanuel TROUVÉ** et **Jocelyn CHANUSSOT**

préparée au sein du **Laboratoire LISTIC**
dans **l'École Doctorale SIE**

Apprentissage Profond pour la Détection Multimodale d'Instabilités Gravitaires Soudaines et Lentes à partir d'Images Satellitaires Bitemporelles

Thèse soutenue publiquement le **01/10/2024**,
devant le jury composé de :

Begüm, DEMIR
Professeure au Technische Universität Berlin, Rapportrice
Germain, FORESTIER
Professeur à l'Université Haute-Alsace, Rapporteur
Yannick, BERTHOUMIEU
Professeur à l'Université de Bordeaux, Président et Examinateur
Sophie, GIFFARD-ROISIN
Chargée de recherche à l'Université Grenoble Alpes, Examinatrice
Loïc, DENIS
Professeur à l'Université de Saint-Étienne, Examinateur
Abdourrahmane, ATTO
Professeur à l'Université Savoie Mont-Blanc, Directeur de thèse
Emmanuel, TROUVÉ,
Professeur à l'Université Savoie Mont-Blanc, Co-directeur de thèse
Jocelyn, CHANUSSOT,
Directeur de recherche à l'INRIA de Grenoble, Co-directeur de thèse


UNIVERSITÉ
SAVOIE
MONT BLANC

Thesis Manuscript

Deep Learning for Multimodal Detection of Sudden and Slow Moving Slope Instabilities on Bitemporal Remote Sensing Images

---

by BRALET Antoine

Supervised by:

ATTO Abdourrahmane M., Professor at Université Savoie Mont-Blanc
TROUVÉ Emmanuel, Professor at Université Savoie Mont-Blanc
CHANUSSOT Jocelyn, Research Director at INRIA Grenoble.

Le plus important est de te réveiller le matin avec le sourire
en pensant à la journée qui t'attend.

Acknowledgements

The entire work done in this thesis is not due to a single PhD student. In fact, it is based on the collaboration of a large number of people who have helped me to focus, innovate, and enjoy these three years.

Necessarily, my first thanks go to the Région Auvergne-Rhône Alpes who financially supported the thesis through the IATOAURA project. Being part of this project I had the opportunity to discover multiple ways to develop artificial intelligence in different domains with the 5 other supported theses.

I would like to thank very much the different jury members who validated my thesis. In particular, I thank the members of the CSI (*Comité de Suivi Individuel*): Alexandre Benoit, Mauro Dalla Mura, Serge Miguet and Razmig Kéchichian for supervising my work yearly and for highlighting the weaknesses of the thesis to help me strengthen the relevance of the research. In addition, regarding my PhD defense, I would like to warmly thank the members of the PhD committee: Begüm Demir and Germain Forestier for their valuable reports, as well as Yannick Berthoumieu, Sophie Giffard-Roisin and Loïc Denis. I greatly appreciate their interest in my work, their remarks and questions during the defense and their availability to evaluate the thesis.

Then, from a scientific point of view, I deeply thank the geomorphologists of EDYTEM and ISTerre who collaborated with me. Thanks to Xavier Bodin, Diego Cusicanqui and Pascal Lacroix, I had the pleasure to discover the world of slope instabilities, to understand them and even to carry out some *in situ* measurements. All the meetings and discussions we had actually helped me to better adapt my work to the challenges inherent to slope instability processes.

Continuing with the scientific contributions, I would like to thank the interns who worked under my supervision and helped to explore new ideas to improve the relevance of the thesis work. Many thanks to Kilian Chantreux for the interferogram generation routine, to Sarah Lantrin for the first tests on explainability, to Trong Nghia Ngo for exploring different DEM-SAR fusion strategies, and to Dhia Turki who was extremely challenged by the change detection task on the Haiti dataset.

From both a scientific and a social point of view, I would like to express my deep and sincere gratitude to all the interns, PhD students and researchers I have met during these last three years. In particular, I would like to thank the members of the LISTIC lab in Annecy where I developed most of my work. I would like to thank each of them individually but I think that the "Acknowledgements" section should be rather short. But I cannot forget the scientific and dynamic discussions in the open spaces, my first hike in the Tournette, the hilarious werewolf games, the movie sessions or even the ridiculous theorems we came up with. I think I have at least one specific moment in my memory with each person from the lab that helped me to make this thesis journey peaceful, and for that I thank you very much.

Special thanks and thoughts go of course to my PhD supervisors. Between the long long long scientific discussions with Abdourrahmane M. Atto, the amount of knowledge and opportunities provided by Emmanuel Trouvé, and the daily support and the encouragement for innovation from Jocelyn Chanussot, I have to admit that I was definitely a lucky PhD student. You all gave me the strength and knowledge to persevere in my work, you supported me in the great and hard times, and you gave me more opportunities to shine than I could have expected. You have definitely helped me to find my way and I don't want to leave it! For these three years, I wanted to send you my special thanks, because these works are not only mine, but yours as well.

Last but not least, I would like to end this page with some French words to my family. Un grand merci à vous tous de m'avoir encouragé à continuer sur cette voie et d'avoir été derrière moi à chacun de mes pas.

Je sais que j'ai été très globe-trotter durant mes études et que je vous ai fait tourner en bourrique, mais je vous rassure, mes études se terminent enfin... mais pas sûr que ça m'empêche de me promener. Grâce à vous j'ai pu comprendre ce que c'était que de vivre de sa passion, je sais désormais ce que cela fait de se lever le matin avec le sourire. Papa, Maman, Gamin, et toutes les personnes qui se reconnaîtront dans ces mots, merci infiniment pour tout ! J'ai nécessairement une pensée émue pour toutes les personnes qui nous ont quittées et qui ont aussi contribué à faire de moi ce que je suis aujourd'hui. Cette année n'a pas été une année des plus joyeuses mais je peux désormais annoncer grâce à vous qu'elle n'est pas tout à fait désastreuse car elle marque le début d'une nouvelle aventure !

En bref et sans images :

Un Immense Merci à Tous !

Abstract

The objective of this thesis is to leverage the wide spatial coverage and the high acquisition frequency of remote sensing images to detect and localize slope instability processes. Two categories of phenomena are targeted - namely sudden landslide failures and slow sliding slope surfaces. While the former can be detected by change detection approaches despite the coarse resolution of remote sensing images (decametric order), the latter are expected to move only of few centimeters between successive acquisitions making them imperceptible. Therefore, in the thesis, different methods have been applied to detect both sudden and slow moving phenomena.

Based on past researches, the use of optical images to detect sudden landslides is more promising than the use of Synthetic Aperture Radar (SAR) images. Nevertheless, landslide collapses are mostly triggered by intense rainfall and the dense cloud cover of such events prevents the use of optical images. Therefore, the first stage of this thesis consists in translating the weather-independent SAR images into optical images to make them exploitable for landslide detection. A novel SAR Distorted Image translator Network (SARDINet) is implemented not only to perform the SAR-to-optical translation but also to account for SAR geometrical distortions caused by the underlying topography. Additional experiments are conducted to improve the relevance of the optical reconstructed patterns such as exploring the best way to fuse the Digital Elevation Model (DEM) with the SAR image as input to the translator, or reconstructing more relevant patterns based on the upsampling strategy used in the decoder. Similarly, the contrast enhancement is addressed by proposing an improved SARDINetV2 architecture and by implementing a Translation-Deblurring (Tra-De) framework based on a post-translation adversarial network.

The second stage of the thesis is to perform a change detection task between a pre-event optical image and a post-event translated SAR image. However, based on the results obtained by a standard Change Detection Network (CDNet), the features reconstructed by the translator network are not characteristic enough of the land-cover to apply a direct change detection task. Therefore, we introduce the Translation-Classification Loss (TCL) to better optimize the training of the translator. The TCL evaluates both the quality of the translated optical images and their ability to be classified by an auxiliary downstream classifier. The translator is then further fine-tuned with a Translation-Change Detection Loss (TCDL), an extension of the TCL for the change detection task, to leverage both the land-cover knowledge of the TCL and the landslide specialization of the TCDL. Based on first experimental results, the application of CDNet to the translation resulting from the TCDL shows better results than the vanilla translations.

The final stage of this thesis focuses on the detection of slow sliding slope surfaces including rock glaciers and slow moving landslides. Due to their displacement of only a few centimeters, they cannot be detected by amplitude correlation techniques on 10m resolution remote sensing images such as Sentinel-1 (SAR) or Sentinel-2 (optical). Therefore, in collaboration with geomorphologists, an annotation routine is established to detect such areas on the phase difference of SAR interferograms (InSAR). It results in the creation of the first InSAR dataset for Slow SLIding areas DEtection (ISSLIDE) which is manually annotated and exploitable with deep learning approaches. In addition to standard deep network architectures, we developed a novel methodology to train a network to be explainable. An Explainability-Constrained claSsifier for Pairing the detection and Localization of slow moving Areas on INterferograms (ECSPLAIN) is implemented. It trains the classifier both for its ability to detect surface motion and to explain itself by localizing the detected phenomena through the generation of a Class Activation Map (CAM). Finally, the classifier is applied to enrich the ISSLIDE dataset by providing both the detected segmentation and the localization scores obtained over the entire available interferograms.

Résumé

L'objectif de cette thèse est d'exploiter la vaste couverture spatiale et la répétitivité des acquisitions des images de télédétection pour localiser des instabilités gravitaires. Deux catégories de phénomènes sont étudiées ici : les glissements de terrain soudains et les mouvements de terrain lents. Si les premiers peuvent être détectés dans des images satellitaires basse résolution (de l'ordre du décamètre) par des approches de détection de changements, le déplacement des seconds est tellement faible (de l'ordre du centimètre) qu'il y est imperceptible. C'est pourquoi, plusieurs méthodologies sont investiguées dans cette thèse pour traiter à la fois les phénomènes soudains et ceux à faible vélocité.

D'après la littérature sur le sujet, la détection de changements appliquée aux glissements de terrain soudains sur des images optiques est plus performante que sur des images radar à synthèse d'ouverture (SAR). Cependant, ces phénomènes étant déclenchés par de fortes précipitations, la couverture nuageuse ne permet souvent pas l'utilisation des images optiques. Ainsi, le premier volet de ces travaux de thèse traite de la traduction d'images SAR, insensibles à la météorologie, en images optiques, pour faciliter la détection de glissements. La création d'un réseau de neurones traducteur d'images radar (SARDINet) a permis de générer des images optiques en limitant les problèmes de distorsions géométriques radar causées par la topographie. Plusieurs expériences ont permis d'améliorer la pertinence des motifs optiques reconstruits telles que fusionner un Modèle Numérique de Terrain (MNT) à une image SAR en entrée du traducteur, ou étudier différentes stratégies de sur-échantillonnage dans le décodeur. Le réhaussement du contraste est aussi amélioré grâce à une nouvelle architecture, SARDINetV2, et une stratégie de Traduction-Défloutage (Tra-De) basée sur un réseau adversaire appliqué sur les traductions.

Le deuxième volet de cette thèse vise à appliquer une tâche de détection de changements entre une image optique pré-évènement et une image SAR traduite post-évènement. Mais, d'après les résultats obtenus par un réseau de détection de changements standard (CDNet), les caractéristiques reconstruites par le traducteur ne sont pas suffisamment spécifiques à la couverture du sol pour permettre une détection de changements. Nous introduisons ainsi une nouvelle fonction de coût, la TCL, pour optimiser le traducteur en évaluant la qualité des images traduites et leur classificabilité avec un classificateur auxiliaire. Le traducteur est ensuite raffiné grâce à une fonction de coût dérivée de la TCL et adaptée à la détection de changements (TCDL) pour bénéficier des acquis sur la couverture du sol obtenus avec la TCL et se spécialiser sur la détection de glissements de terrain avec la TCDL. Les premières expériences montrent qu'appliquer un CDNet sur les traductions résultantes d'un entraînement avec la TCDL est plus performant que sur les traductions initiales.

Le dernier volet de cette thèse traite de la détection de mouvements gravitaires lents (*e.g.* glaciers rocheux et glissements de terrain lents). Les déplacements recherchés étant centimétriques, la corrélation d'amplitude d'images satellitaires de résolution décamétrique (*e.g.* Sentinel-1 (SAR) et Sentinel-2 (optique)) ne peut pas être envisagée. Ainsi, une routine d'annotation a été proposée en collaboration avec des géomorphologues pour détecter ces mouvements par différence de phase dans des interférogrammes d'images SAR (InSAR). Cette routine a permis d'annoter manuellement un nouveau jeu de données adapté à des modèles d'apprentissage automatique visant à détecter des mouvements lents sur des interférogrammes (ISSLIDE). En plus d'entraîner des modèles de segmentation standards, une nouvelle stratégie d'apprentissage incitant le réseau de neurones à être explicable est proposée. Un classificateur dont l'explicabilité est contrainte est ainsi implémenté pour détecter et localiser ces mouvements lents sur des interférogrammes (EC-SPLAIN). Le classificateur est entraîné à la fois sur sa faculté à détecter les mouvements en surface, et à se justifier en localisant le phénomène détecté à l'aide d'une carte d'activation (CAM). Ce modèle est utilisé pour enrichir ISSLIDE avec la segmentation et la carte de localisation de chaque interférogramme.

Contents

Abstract **7**

Résumé **9**

List of publications **15**

List of abbreviations **17**

1 Introduction **21**
1.1 Appetizer . . . 21
1.2 Remote Sensing Imagery . . . 21
1.3 Slope Instability Threats . . . 22
1.4 Deep Learning . . . 24
1.5 Thesis Objectives . . . 25
1.5.1 Overcoming Image Challenges with SAR-to-optical Translation . . . 26
1.5.2 Detecting Landslides Based on Translated Images . . . 26
1.5.3 Explaining Slow Moving Slope Instability Detections . . . 27
1.6 Thesis Overview . . . 28

2 Geoscience and Remote Sensing Context **29**
2.1 Introduction . . . 29
2.2 Slope Instability Processes . . . 29
2.2.1 Phenomena of Interest . . . 29
2.2.2 Monitoring Stages . . . 33
2.2.3 Standard Approaches . . . 34
2.3 Remote Sensing Images . . . 35
2.3.1 Optical Images . . . 36
2.3.2 Synthetic Aperture Radar Images . . . 38
2.4 Radar Interferometry . . . 42
2.4.1 Exploring SAR Phase . . . 42
2.4.2 Generating Interferograms . . . 44
2.4.3 Applying to Slope Instabilities . . . 45
2.5 Conclusion . . . 46

3 Deep Learning Context **47**
3.1 Introduction . . . 47
3.2 Deep Learning for Image Processing . . . 47
3.2.1 Overall Context . . . 47
3.2.2 From U-Net to Transformers . . . 50
3.2.3 Image-to-Image Applications . . . 54
3.2.4 Performance Evaluation . . . 55

3.3 Modality Translation . . . 58
3.3.1 From Computer Vision to Remote Sensing . . . 58
3.3.2 Supervised Adversarial Networks . . . 61
3.3.3 Unsupervised Adversarial Networks . . . 63
3.3.4 Towards Stable and Reliable Translations . . . 64
3.3.5 Conclusion . . . 66
3.4 Class Identification . . . 66
3.4.1 Classification . . . 67
3.4.2 Segmentation . . . 70
3.4.3 Change Detection . . . 72
3.4.4 Movement Detection . . . 74
3.5 Neural Networks Explainability . . . 75
3.5.1 Class Activation Maps . . . 75
3.5.2 CAM-based Localization . . . 79
3.5.3 Right for the Right Reasons . . . 80
3.6 Conclusion . . . 81

4 SAR-to-Optical Modality Translation 83
4.1 Introduction . . . 83
4.2 Materials . . . 83
4.2.1 SpaceNet 6 - Urban Distorted Context . . . 83
4.2.2 BigEarthNet-MM - Rural and Annotated Dataset . . . 84
4.2.3 Mountain Dataset - Typical Slope Instability Regions . . . 85
4.3 Distortion Robust Modality Translator . . . 86
4.3.1 SARDINet Architecture . . . 86
4.3.2 Results . . . 87
4.3.3 SARDINet Outcomes . . . 89
4.4 Architecture Improvements . . . 89
4.4.1 Fusion Strategy . . . 89
4.4.2 Decoding Strategy . . . 94
4.5 Contrast Enhancement . . . 99
4.5.1 SARDINetV2 . . . 99
4.5.2 Adversarial Translations . . . 102
4.5.3 Post-translation Deblurring . . . 103
4.5.4 Translation-Deblurring Framework . . . 105
4.5.5 Contrast Enhancement Outcomes . . . 108
4.6 Conclusion . . . 109

5 Sudden Landslides Detection 111
5.1 Introduction . . . 111
5.2 Materials . . . 111
5.3 Translation-based Landslide Detection . . . 113
5.3.1 Training Strategies . . . 113
5.3.2 SARDINetV2 Fine-tuning . . . 114
5.3.3 Change Detection Architectures . . . 114
5.3.4 Landslides Detection Results . . . 116
5.3.5 Translation-based Change Detection Outcomes . . . 119
5.4 Classification-Guided Modality Translation . . . 120
5.4.1 Translation-Classification Loss . . . 120
5.4.2 Backbone Translators . . . 122
5.4.3 Translation Results . . . 123
5.4.4 TCL Outcomes . . . 128

5.5 Change Detection based Modality Translation . . . 129
5.5.1 Translation-Change Detection Loss . . . 129
5.5.2 SARDINetV2 Fine-tuning with TCDL . . . 130
5.5.3 Change Detection Results . . . 132
5.6 Conclusion . . . 133

6 Slow Moving Area Detection **137**
6.1 Introduction . . . 137
6.2 Materials . . . 137
6.2.1 Interferogram Generation . . . 137
6.2.2 Annotation Process . . . 139
6.2.3 ISSLIDE Dataset . . . 141
6.3 Standard Deep Segmentation Networks . . . 143
6.3.1 Standard Architectures . . . 143
6.3.2 Segmentation Results . . . 144
6.3.3 InSAR Segmentation Outcomes . . . 146
6.4 Localization through Learnt Explainable Detection . . . 147
6.4.1 Explainable Classification or Segmentation . . . 147
6.4.2 ECSPLAIN Framework . . . 149
6.4.3 Slow Moving Area Localization Results . . . 150
6.5 ECSPLAINing ISSLIDE . . . 157
6.5.1 Detection Pipeline . . . 157
6.5.2 Detected Instabilities . . . 158
6.6 Conclusion . . . 160

7 Conclusion and Perspectives **163**
7.1 Modality Translation . . . 163
7.2 Bitemporal Change Detection . . . 164
7.3 Interferogram Inventorying . . . 165
7.4 Gates Are Opened! . . . 166
7.4.1 Towards more Relevant Translations . . . 166
7.4.2 Towards more SAR-optical Feature Comparisons . . . 167
7.4.3 Towards more Geophysically-based Detections . . . 168
7.5 One Last Word . . . 169

List of publications

Conferences

Published

- *Deep Learning of Radiometrical and Geometrical Sar Distorsions for Image Modality translations*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E. (2022, October). In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 1766-1770). Bordeaux, France, IEEE. DOI: 10.1109/ICIP46576.-2022.9897713.

  Available on HAL: https://hal.science/hal-03844839v1

- *Impact de la stratégie de décodage sur la traduction de modalité radar-optique d'images de télédétection*, Bralet, A., Atto, A., Chanussot, J., Trouvé, E., number 2023-1309, pages p. 929–932, Grenoble. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

  Available on HAL: https://hal.science/hal-04191879v1

- *DEM-assisted neural network for SAR-to-optical image translation*, Bralet, A., Ngo, T. N., Trouvé, E., Chanussot, J., Atto, A. M., IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2024, Athens, Greece.

  Available on HAL: https://hal.science/hal-04720021

Journals

Published

- *ISSLIDE: A new InSAR dataset for Slow SLIding area DEtection with machine learning*, Bralet, A., Trouvé, E., Chanussot, J., Atto, A. M., in IEEE Geoscience and Remote Sensing Letters, doi: 10.1109/LGRS.2024.3365299.

  Available on HAL: https://hal.science/hal-04482461

Submitted

- *Translation-classification loss for SAR image understanding with deep learning*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E., submitted to Computer Vision and Image Understanding (CVIU), Elsevier submitted on April 19, 2024.

- *ECSPLAIN: Explainability Constrained-claSsifier for Pairing detection and Localization of moving Areas on INterferograms*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E., submitted to IEEE Transactions on Geoscience and Remote Sensing (TGRS) submitted on June 10, 2024.

Datasets

- *SAR-DEM-Optical Mountainous Dataset for Distortion Management*, Bralet, A., Ngo, T. N., Trouvé, E., Chanussot, J., Atto, A. M., January 3, 2024, IEEE Dataport, doi: https://dx.doi.org/10.21227/16ce-ce92.

  Available on IEEE Dataport:
  https://ieee-dataport.org/documents/sar-dem-optical-mountainous-dataset-distortion-management

- *Multimodal Remote Sensing Dataset for Landslide Change Detection in Haiti*, Bralet, A., Trouvé, E., Chanussot, J., Atto, A. M., July 14, 2024, IEEE Dataport, doi: https://dx.doi.org/10.21227/4heb-7h07.

  Available on IEEE Dataport:
  https://ieee-dataport.org/documents/multimodal-remote-sensing-dataset-landslide-change-detection-haiti

- *ISSLIDE: InSAR dataset for Slow SLIding area DEtection with machine learning*, Bralet, A., Trouvé, E., Chanussot, J., Atto, A. M., September 22, 2023, IEEE Dataport, doi: https://dx.doi.org/10.21227/dhxt-5g91.

  Available on IEEE Dataport:
  https://ieee-dataport.org/documents/isslide-insar-dataset-slow-sliding-area-detection-machine-learning

List of abbreviations

- Acc. : Accuracy
- ACoL : Adversarial Complementary Learning
- Adv-SARDINet : Adversarial SARDINet
- AUC : Area Under the ROC Curve

- BCE : Binary Cross Entropy
- BN : Batch Normalization

- CAM : Class Activation Map
- CBAM : Convolutional Block Attention Module
- CBN : Conditioned Batch Normalization
- CDNet : Change Detection Network
- cGAN : Conditional Generative Adversarial Network
- CNN : Convolutional Neural Network
- Conv1D/Conv2D : 1D/2D Convolutional layer
- CRESLIDE : Coupled Routing and Excess Storage and Slope-Infiltration-Distributed Equilibrium model
- CRF : Conditional Random Field

- DEM : Digital Elevation Model
- DFT : Discrete Fourier Transform

- ECSPLAIN : Explainability-Constrained claSSifier for Pairing detection and Localization of moving Areas on INterferograms
- EO : Earth Observation
- ESA : European Space Agency

- F1-S : F1-Score
- FC : Fully Connected layer
- FCN : Fully Convolutional Network
- FID : Fréchet Inception Distance
- FLOP : Floating Point Operation

- GAN : Generative Adversarial Network
- GAP : Global Average Pooling
- GCGAN : Geometry-Consistent Generative Adversarial Network
- GPU : Graphics Processing Unit
- GRD : Ground Range Detected
- GRU : Gated Recurrent Unit

- HH (resp. HV) : Horizontally polarized emission, Horizontally (resp. Vertically) polarized reception
- HD : Hausdorff Distance

- IGN : Institut National de l'information Géographique et forestière
- InSAR : Interferometric SAR
- ISSLIDE : InSAR dataset for Slow SLIding area DEtection
- IW : Interferometric Wide swath mode

- LOS : Line Of Sight
- LSTM : Long Short Term Memory

- MAE : Mean Absolute Error
- MLP : Multi-Layer Perceptron
- MSE : Mean Square Error
- mIoU : mean Intersection over Union

- NASA : National Aeronautics and Space Administration
- NICE-GAN : No-Independent-Component-for-Encoding Generative Adversarial Network
- NIR : Near Infra-Red
- NLP : Natural Language Processing

- PCA : Principal Component Analysis
- PINN : Physics Informed Neural Network
- Prec. : Precision
- PS-InSAR : Point-Scatterer InSAR
- PSNR : Peak Signal to Noise Ratio

- R2BU : Ready to Be Used dataset
- Rec. : Recall
- ReLU : Rectified Linear Unit
- RGB : Red Green Blue
- RMSE : Root Mean Square Error
- RNN : Recurrent Neural Network

- ROC : Receiver Operating Characteristic

- SAR : Synthetic Aperture Radar
- SARDINet(V2) : SAR Distorted Image translator Network (improved version)
- SBAS : Short Baseline Subsets
- SepConv : Separable Convolution
- SGD : Stochastic Gradient Descent
- SHALSTAB : SHAllow Landsliding STABility model
- SINMAP : Stability INdex MAPping
- SLC : Single Look Complex
- SNAP : Sentinel Application Platform
- SOTA : State Of The Art
- SRM : Style Recalibration Module
- SRTM : Shuttle Radar Topography Mission
- SSIM : Structural Similarity Index Measure
- SVD : Singular Value Decomposition

- TCDL : Translation-Change Detection Loss
- TCL : Translation-Classification Loss
- TCN : Temporal Convolutional Network
- TiVaSS : Time Variant Slope Stability model
- TOPSAR : Terrain Observation with Progressive Scans SAR
- Tra-De : Translation-Deblurring framework
- TRIGRS : Transient Rainfall Infiltration and Grid-based Regional Slope-stability model

- VAE : Variational AutoEncoder
- VHR : Very High Resolution
- ViT : Visual Transformer
- VV (resp. VH) : Vertically polarized emission, Vertically (resp. Horizontally) polarized reception

- WGS84 : World Geodetic System 1984

Chapter 1

Introduction

1.1 Appetizer *- "What do you do? "...*

... "Basically, I ask my computer to look for slope instability-related phenomena on remote sensing images". Such an answer may be suitable for vulgarization purposes but still there is nothing *basic* in this task. SARDINet for modality translation, Tra-De for visualization, TCL and TCDL for guidance, ISSLIDE for detection, or ECSPLAIN for understanding, so many acronyms that may sound awkward but play their part in localizing slope instabilities in mountainous regions. In this thesis, we propose to familiarize the reader with all these novel concepts and more.

Starting from the beginning, three main topics are at the core of this thesis: remote sensing imagery, slope instabilities, and deep learning algorithms. The actual objective is broader than the straightforward implementation of novel deep learning algorithms capable of detecting instabilities on remote sensing images. The thesis mainly focuses on addressing the various challenges and expectations related to each domain - *i.e.* computer science, image processing and geosciences - in order to develop the most appropriate and adapted methodologies. In particular, memory and time consumption, image modality, temporality and resolution, and geomorphological evidence and explanation are some of the issues addressed in this work. The following sections give more insight into the overall challenges faced during the thesis with respect to each independent domain, before giving an overview of our main contributions.

1.2 Remote Sensing Imagery *- "An image is an image"...*

..., so is the standard idea resulting from dealing with dogs and cats images for too long. However, a pixel is not always *only* a pixel.

In remote sensing imagery, a pixel is directly related to physical information. In Earth Observation (EO), each pixel of a given image refers to the actual physical properties of the underlying terrain. In particular, the wide variety of sensors on board satellites can produce different types of images. These so-called image modalities do not reflect the same physical properties of the underlying terrain. Although satellite optical sensors can acquire Red, Green and Blue (RGB) and Near Infra-Red (NIR) channels to capture the chromatic properties of the ground, as standard computer vision images would do, they are actually *multispectral* sensors capable of acquiring up to tens of spectral optical bands. Yet each spectral band does not reflect the same property as the others and may not even be acquired at the same resolution. Therefore, can the standard three-channel-based methods of computer vision be extended to multi-spectral and multi-resolution images?

From another sensor perspective, the acquisition process consists of sensing the structural properties of the underlying terrain. Indeed, when considering Synthetic Aperture Radar (SAR) images, a single spectral bandwidth is captured. The peculiarity of this sensor lies in its ability to discriminate between the polariza-

tions of the sensing wave. Based on the polarization of the signal backscattered by the objects on the ground, further processing can be applied to understand their structures. Depending on the bandwidth, the electromagnetic waves can even penetrate some media such as clouds, forest canopy or snow cover. In addition, SAR sensors are capable of emitting their own electromagnetic waves so that the resulting backscattered image can be registered in both amplitude and phase, assigning a complex value to each pixel of the image. Another issue associated with SAR imagery is its sensitivity to the slope of the underlying terrain. Indeed, steep slopes in mountainous regions affect SAR images by creating critical geometrical distortions, such as layovers and shadows, which make the image much less interpretable and processable. Given these characteristics, are standard optical approaches transferable to this structural and potentially distorted information acquired with complex values?

Apart from these modality issues, one of the strengths of remote sensing images is the repeatability of their acquisitions. Such frequent acquisitions allow the application of multi-temporal approaches to better understand the evolution of the region of interest. Such a temporal aspect could be associated with video processing in computer vision. Yet, in addition to the modality issues mentioned above, the notion of time is also different in remote sensing. Except the geostationary satellites which are constantly focused on the same area, most satellites have a periodicity ranging from a few days to several weeks. Therefore, the sequence of images is a coarse-sampled video, which is why they are called *time series* in the following. This property further complicates the application of standard methods which assume continuity from frame to frame.

Indeed, from an image point of view, the external acquisition conditions can change drastically from one image to another. For instance, the multispectral images can be significantly affected by the solar exposure of the region of interest. This breaks the temporal continuity of pixel values. Even more disturbing is the presence of clouds in some images. Indeed, they can hide the underlying terrain and render the optical images useless under certain weather conditions. Such sensitivity makes time series with missing values or with variable temporal delay a common issue in the field. In addition, although SAR images are weather independent, they are much less interpretable by non-experts and have poorer spectral information. Such considerations bring multimodality, *i.e.* the combined use of different modalities, and in particular modality translation, to the forefront when dealing with image time series.

Another aspect that limits the use of video-based methods is the possible changes that can occur between two acquisitions. The continuity of certain regions can be broken by sudden changes in the landscape, such as the demolition of buildings, natural disasters (*e.g.* landslides, floods, forest fires) or crop harvesting. However, such breaks in continuity can also be the sought-after information, as they allow the implementation of change detection approaches to better understand the events that have occurred. A typical example is inventorying the location and extent of natural phenomena to enable early response. Sudden changes can also occur due to meteorological conditions and/or seasonal variations. In particular, if a satellite takes an image once a month, the seasonality can significantly affect the interpretation and comparison between two consecutive images. Similarly, a sudden snowfall can prevent multispectral sensors from capturing the ground characteristics and disturb SAR sensors since their ability to penetrate snow depends on its wetness.

Therefore, a remote sensing image is not *just* an image but brings much more information about the ground than 8-bit encoded RGB values. This thesis focuses on understanding the strengths and weaknesses of each sensor modality. In this way, the idea is to better address the development of algorithms capable of accurately detecting slope instabilities by enhancing the strengths of multimodal remote sensing images.

1.3 Slope Instability Threats *- "When a donkey slides into your living room"...*

... This absurd sentence was used to attract the attention of the audience during the "My thesis in 3 minutes" contest[1] by extrapolating the possible consequences of a slope instability failure. Far from being a prediction, a similar event occurred in the city of Canoas in Brazil on May 8, 2024[2]: a horse was found

[1] https://www.youtube.com/watch?v=zgBGXAcv084, last visited on June 26, 2024 (French version only).
[2] https://www.bbc.com/news/av/world-latin-america-68981957, last visited on June 26, 2024.

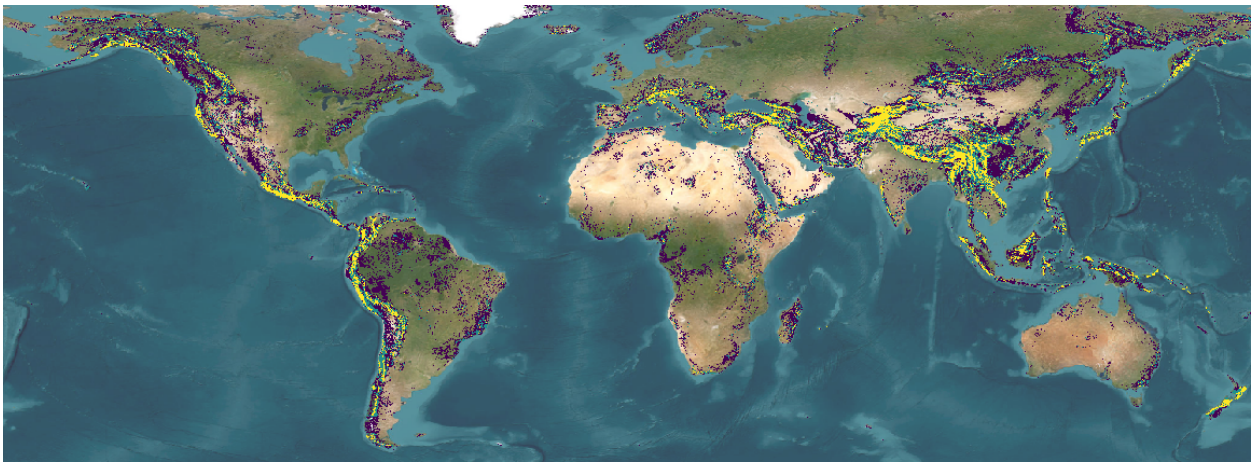


Figure 1.1: Global landslide hazard map provided by World Bank Group[4]. Landslide susceptibility is categorized into 3 classes: low (purple), medium (green) and high (yellow) susceptibility.

standing on a roof to escape flooding and landslides caused by heavy rainfall. Although this event may make one smile, it also brings back the serious threats posed by natural hazards, as more than a hundred people died during this disaster.

In this thesis, slope instabilities are chosen as the main object of study, including both sudden and slow moving landslides, but also rock glaciers, which are increasingly considered as a threat. Indeed, after the first reported collapse which happened in the Bérard rock glacier in the southern French Alps in 2006 [Bodin et al., 2017], a series of rock glacier events occured, for instance two of them in the South Tyrol in the Italian Alps [Kofler et al., 2021] and the Lou rock glacier in the French Alps [Marcer et al., 2020]. Conversely, landslides are already known to be a threat to people and infrastructures. Typical examples include the La Clapière landslide [Fruneau et al., 1996, Delacourt et al., 2004] and the Harmalières landslides [Lacroix et al., 2018] in France, landslides in the northern Apennines in Italy [Bayer et al., 2017] or those near a reservoir in Granada [Mirmazloumi et al., 2022]. Yet, since the latter are slow moving landslides, they can be monitored by the mean of early warning systems to prevent catastrophic consequences. The most dangerous phenomena are the sudden landslides because their collapse is less predictable which makes them the cause of high costs in terms of human lives and damage to buildings. A wide variety of such landslides can be found in Dave Petley's blog[3], including the Mocoa landslide in Colombia [Prada-Sarmiento et al., 2019] or the Xinmo landslide in China [Intrieri et al., 2017].

Landslide collapse is a real and growing concern as climate change may lead to more intense events. Indeed, the main triggering factors for landslide collapse are both seismic hazards and intense rainfall events. In particular, the increase in the intensity of rainfall events, including typhoons, storms and monsoons, can be the cause of more significant landslide failures due to the magnitude of the precipitation that fluidizes the already fragile subsurface layers. The World Bank Group together with the Global Facility for Disaster Reduction and Recovery (GFDRR), and Arup published a report on landslides events from 1980 to 2018[4]. Among the conclusions of the report, the authors found that there was a significant increase in the number of rainfall-induced landslide events reported between the two dates, from 350,000 in 1980 to 625,000 in 2018. By averaging the number of landslides over the 38 years of the study, a total estimate of 400,000 rainfall-induced landslides versus 140,000 earthquake-induced landslides was calculated. Figure 1.1 gives the most landslide-prone areas in the world according to the report. Based on this figure, the mountainous regions are obviously the most vulnerable including the Alps (Europe), the Andes (South America), the Rocky

[3]https://blogs.agu.org/landslideblog/, moved to https://eos.org/thelandslideblog, last visited on June 26, 2024.

[4]https://datacatalogfiles.worldbank.org/ddh-published/0037584/DR0045411/global-landslide-hazard-map-report.pdf?versionId=2023-01-18T20:43:00.6156776Z, last visited on June 26, 2024.

Mountains (North America) and the Himalayas (Asia). In particular, the United States of America, China, India and the Kyrgyz Republic are identified as having the most frequent events during the study period. Even if not all landslide hazards turn into disasters, their frequency and susceptible threats forces several countries to make an estimate of the dangers such as France[5] and Sri Lanka[6] and plan pre- and post-event actions to better anticipate, mitigate and respond to each threat.

One of the most critical issues regarding slope instability is the diversity of these phenomena. Although the result is the same, *i.e.* a catastrophic collapse threatening lifes and infrastructures, and the causes are mainly rainfall or earthquakes, the underlying physical processes can be very different. There is still no global slope instability model capable of accurately predicting the behavior of all instabilities and forecasting their collapse. In addition, slow moving instabilities can move from several millimeters to a few meters per year, making them even more difficult to identify and monitor using open access satellite images with decametric resolution. Nevertheless, even though these phenomena can be variable in the subsurface, their surface characteristics are better known. From an identification point of view, geomorphologists and geoscientists rely on very specific features. In particular, landslides are mainly covered by scarps and slide along a shear surface while rock glaciers can be identified by the presence of ridges and furrows. The consideration of such geomorphological parameters in the explanation of the predictions obtained by a given method is crucial to ensure accurate identifications and the trustworthiness of the algorithms.

1.4 Deep Learning

- "A learning algorithm is like a stubborn child"...

... and like a child it needs to learn, make mistakes and most importantly, be guided to acquire knowledge. Indeed, deep learning approaches are increasingly used because of their strong generalization ability, easy vanilla training (*i.e.* without strong constraints and applied to standard data) and fast inference computation. They have been used for a wide range of applications from computer vision to self-driving cars, from medical imaging to art restoration, from natural language processing to music and video creation, but also in remote sensing and geosciences. However, such rapid development raises many questions and issues with regard to certain applications.

The first important issue is the availability of data for the training step. To ensure accurate and efficient model training, there is no minimum data requirement: the more, the better. A typical example is the ImageNet dataset [Deng et al., 2009] which contains more than 14 million images associated with a label corresponding to the class they belong. This last aspect is one of the most limiting issues: the availability of labels, or more generally, of the ground truth associated to each sample. Indeed, as already mentioned, the network is prone to make mistakes during the training phase. Therefore, to avoid such errors, accurate annotations should also be provided to ensure the correct *supervision* of the training and improve the performance of the network. If some *unsupervised* approaches are already made to compensate for the lack of such annotations, they are generally less powerful to solve complex tasks than their supervised counterparts. Therefore, before implementing any deep learning algorithm, considerable effort is required to ensure the relevance of the available ground truth. This aspect is even more important in remote sensing and geosciences, where the data are more challenging and the phenomena of interest can be complex. Therefore, more than ensuring the availability of annotations associated with the data samples, a deep understanding of both the source data and the target task is required to appropriately adapt the design of deep neural networks. Such an understanding will then facilitate the development of efficient, reliable and generalizable models.

Generalizability itself is the second thorny issue. This property is critical because it ensures the applicability of the developed approach to new samples not seen by the network. Typically, a neural network

[5]https://www.orcae-auvergne-rhone-alpes.fr/fileadmin/user_upload/mediatheque/ORCAE/Documents/Publications/ORECC-EtatConnaissances_MouvementsTerrains-Aout2017.pdf, last visited on June 26, 2024 (French version)

[6]https://www.iwmi.cgiar.org/Publications/Other/PDF/development_of_an_anticipatory_action_plan_for_floods_and_landslide_hazard_in_sri_lanka.pdf, last visited on June 26, 2024

trained to detect slope instabilities in the Alps may not be able to detect them in the Andes or the Himalayas. Among the various reasons, two are common in practice: domain shift and overfitting. Domain shift is related to the distribution of the training data compared to the inference data. The vegetation, the steepness of the slopes or their orientation are typical examples of differences from one mountain massif to another that can cause domain shift. Since the network has never seen such configurations, generalizing its knowledge to these new data can lead to a drop in performance, which should be avoided. Conversely, the problem of overfitting results from inappropriate training. Basically, it consists in the specialization of the neural network to process the training dataset and not other samples. Therefore, the model cannot be accurately applied at inference time except to data that are already known. Different strategies can be adopted to mitigate the effects of both domain shift and overfitting issues on neural networks which requires a special attention in identifying their causes whether due to the data or to the network itself.

The third training characteristic to consider is related to the networks themselves. From an architectural point of view, two different trends are raising the field. On the one hand, to ensure greater robustness and generalizability, architectures are becoming more complex by adding more learnable parameters, including attention modules, deepening the architecture by cascading more and more layers, and increasing the size of these layers to extract even more features. Such architectural changes can lead to extremely heavy models which challenges their portability in limited memory devices. In addition, such architectures are more prone to overfitting when trained on small datasets. They require not only a large memory capacity but also a huge training dataset which can take weeks to months to train the network. On the other hand, lightweight networks are also developed to limit the latter issues. The main consequence of such a reduction of learnable parameters is a susceptible drop in performance when applied to complex tasks since the network cannot learn as many variable features as the largest networks. For this reason, significant efforts should be made to identify the limitations of each network in order to better balance its performance with respect to its computational expectations.

Last but not least, regardless of their size, generalizability and input data, neural networks are widely considered to be *black boxes*. This term refers to an algorithm that is poorly understandable. In particular, even if a prediction made by the algorithm is true, there are no clues to confirm whether it is the expected result or not. Indeed, at inference time, all network parameters are frozen and the network can make a decision from any input. Yet, there is no explanation of the reasons that lead to a given decision instead of another. Such a lack of explainability is of great concern in the geosciences since decisions, and especially responses to susceptible threats, cannot be made without strong evidence. The relationship between the input data and the provided decision should be explicit and understandable in order to be trusted. Despite the fact that a number of works develop explanation methods, most of them try to guess the behavior of the network. However, as the catch phrase of the Section says: neural networks are children. They cannot make correct decisions if they are misguided or not guided at all. Therefore, more than trying to understand the decisions, an important issue is to ensure that these decisions are based on accurate and reliable cues.

Although these challenges are inherent to deep learning approaches, they are strengthened when combined with the issues of remote sensing imagery and slope instability phenomena. While the former are readily available to create large datasets, few of them are already accurately annotated. In addition remote sensing images are already heavy images that can require up to several gigabytes of memory, making the burden on a processing device even heavier when using large neural networks. From a slope instability perspective, the variability of the phenomena can cause troubles in terms of network generalizability while experts in the field are particularly demanding in terms of decision understanding.

1.5 Thesis Objectives

- "Studious algorithms is my way of keeping the donkeys away"...

... This is the main issue raised by the above discussions. Indeed, early detection of slope instability could help to mitigate the damage caused by natural hazards. The key challenge is to make deep learning algorithms smarter than they are, in particular by ensuring accurate guidance for more reliable and trustworthy

decisions. To achieve these objectives, different challenges and research questions should be considered.

1.5.1 Overcoming Image Challenges with SAR-to-optical Translation

Before addressing the target task, a first challenge is data availability. Indeed, the detection of slope instability is easier on optical images but these images lack temporal consistency because they are affected by the presence of clouds. In addition, as reported by the World Bank Group, landslides are mainly triggered by rainfall events, which by definition implies the presence of dense cloud cover. An alternative could be the use of SAR images to replace the cloudy optical images. Yet, two problems arise from such an approach: performance degradation and image interpretability. Even though SAR images are weather-independent and reveal the structural properties of the underlying terrain, they are difficult to understand by non-experts. In addition, their use for classification, segmentation and change detection tasks is more challenging than optical images and may miss real threats due to the lack of spectral information. Therefore, the first challenge to be addressed is:

How to make SAR images understandable and useful for slope instability detection?

Furthermore, by definition, slope instabilities occur mostly in regions with steep slopes such as mountainous or hilly areas, as illustrated by the World Bank Group's landslide susceptibility map in Figure 1.1. The presence of steep slopes has a significant impact on SAR acquisitions, creating geometrical distortions that prevent any interpretation in the so-called layover and shadow areas. Therefore a second challenge should be considered here:

How to limit the effects of geometrical distortions to improve the interpretability of SAR images?

The first set of contributions proposed in this thesis aims at solving the latter two challenges. They are addressed through the implementation of novel *SAR-to-optical modality translation* approaches and in particular with a SAR Distorted Image translator Network (SARDINet). Indeed, the translation of SAR structural properties into radiometric properties would generate a false optical image that could replace a cloudy true optical image. Taking into account the geometrical distortions in the training of the network through external guidance could even help to limit the impact of the distortions on the optical reconstructions, making them more interesting for downstream applications, such as slope instability detection.

1.5.2 Detecting Landslides Based on Translated Images

Focusing on the fast detection of landslide collapse, the idea is to leverage the proposed translation capabilities to apply a downstream change detection task. Yet, additional concerns should be considered. Indeed, using translation as a prior to another task may lead to a domain shift. Since the task is performed by a downstream neural network, the distribution and features extracted when trained on real optical images may differ from those of the images generated by the translation network. A simple solution would be to retrain the downstream network only on the translated images. But this strategy would not take advantage of the true optical images, which necessarily contain more accurate features with respect to the task of interest. In addition, SAR-to-optical image translations are not originally trained to reconstruct landslide-like patterns since these events are rare in the world and absent in most datasets. Thus, the translator should transfer the knowledge acquired on clean images to a novel context with specific slope instability patterns. This leads to a third challenge:

How to make the SAR-to-optical translation relevant to a downstream task, specifically landslide collapse detection?

Even more restrictive is the fast detection of landslide collapse through change detection approaches. Such a task is based on the comparison between a pre-event and a post-event image. However, if the translated image does not have similar features nor similar brightness and colors, this comparison would lead to a drop

in performance. For this reason, not only should the reconstructed patterns be relevant to the downstream task, but they should also be comparable to another true optical image acquired before the event. Thus, the fourth challenge of this thesis can be stated as:

How to make the translations comparable to pre-event images?

These two challenges are at the core of the second set of experiments and contributions of this thesis. Different strategies can be implemented to address both issues. In particular, one can improve the change detection capabilities of the downstream network to mitigate the domain shift through domain adaptation approaches or focus on more relevant SAR-to-optical translation. The second approach is considered in this thesis, since improving the translation capabilities would allow a better understanding of the SAR image making it even more interpretable for non-experts. For this purpose, novel strategies based on a pre-trained auxiliary network are proposed. This network aims to evaluate the translation performance with respect to the desired downstream task such as classification or change detection. Two joint losses are proposed: the Translation-Classification Loss (TCL) for more relevant land-cover pattern reconstructions and the Translation-Change Detection Loss (TCDL) to focus on slope instability detection. The latter aims to mimic the pre-event image in unaffected areas while reconstructing landslide-like patterns in damaged areas.

1.5.3 Explaining Slow Moving Slope Instability Detections

Although change detection approaches can be applied to sudden landslide collapses, slow moving phenomena are less perceptible on most optical and SAR images. Indeed, the velocity of these movements can be so small that they are invisible even after a year on images with metric to decametric resolution. A solution to estimate the localization of these instabilities is an additional technique to be applied to SAR images. Indeed, SAR interferometry (InSAR) can generate interferograms to exploit the phase of the SAR images and reveal movements up to centimeters. Nevertheless, if SAR images are difficult to understand, interferograms are even more challenging because the pixel value is not related to structural or radiometric features. In addition, a significant number of counterfactual patterns due to external disturbances can show similar behavior as moving areas on these interferograms. All these peculiarities associated with these SAR-derived images present an additional challenge:

How to detect slow moving areas on SAR interferograms?

This particular difficulty adds to the distrust of decisions made by neural networks. Indeed, not only is the algorithm poorly explainable but the target task is also full of tricky patterns that can even mislead a human. This increases the expectations in terms of explainability of the decisions. In addition, slow moving areas cannot be strictly defined because they are not adjacent to an actual stable surface. Indeed, a sharp boundary cannot be drawn because the central motion necessarily generates displacement in its neighborhood. Therefore the challenge for the implemented neural network is twofold:

How to ensure reliable and explainable detection of phenomena with uncertain boundaries?

These are two additional challenges that are addressed by the final set of contributions of the thesis. In particular, the detection of slope instabilities on interferograms is studied in detail in collaboration with geomorphologists in order to elaborate precise guidelines. These guidelines aim first to better understand the behavior of the instabilities, and second to identify the differences between these phenomena and counterfactual patterns on interferograms. They are applied to the creation of a novel Interferogram dataset for Slow SLIding area DEtection (ISSLIDE) which is manually annotated and calibrated for deep learning strategies. However, incorporating these guidelines as prior knowledge into deep learning algorithms is more challenging because it requires the network to be able to accurately explain itself. This self-explainability is implemented by training the network to align its explanation with the expectations arising from the previously established guidelines. In particular, an Explainability-Constrained claSsifier for Pairing detection and Localization of moving Areas on INterferograms (ECSPLAIN) is developed. This constraint forces the classifier to make its decisions based on the manual annotations provided in ISSLIDE. In addition, the computation of the explanation of the classifier generates a localization map that avoids the aforementioned sharp detection issue. It results in a shallow network capable of outputting a pair of detection score and localization map.

1.6 Thesis Overview

According to the aforementioned issues, objectives and solutions, the rest of the thesis is divided in five different chapters. The following Chapter 2 gives more insight into both slope instability processes and remote sensing imagery. Without delving deeply into subsurface processes, geomorphological considerations or geoscientific debates, it first introduces the different types of slope instability considered in the thesis as well as the standard physical models used to predict their behavior, in particular their collapse. Then, multispectral and SAR imagery are described in detail, focusing on their respective advantages and drawbacks. A brief review of their use in slope instability monitoring helps to better understand the pros and cons of coupling remote sensing and slope instability domains. SAR interferometry is then presented as a powerful tool for monitoring slow moving phenomena despite the different issues associated with its interpretation and disturbances.

Deep learning algorithms are the topic of Chapter 3. Since this thesis uses a number of strategies, concepts and notions in deep learning, they need to be clearly defined to avoid confusion with the various vocabulary available in the literature. Therefore, a comprehensive yet brief introduction to deep learning is first given to highlight the key concepts of these algorithms when applied to images. Then, the focus is on detailing standard and remote sensing modality translation strategies from the literature to better understand the first set of contributions of this thesis. The terminology and an extended literature on the identification of classes within remote sensing images is then provided. The objective is to clearly differentiate classification, segmentation, change detection and localization strategies. At the end of the chapter, further insight into the explainability of neural networks are given to better figure out the need for explainability but also the drawbacks of current approaches. Special focus is given to Class Activation Maps (CAMs) which are widely used for localization purposes.

The first set of contributions targeting SAR-to-optical modality translation is then given in Chapter 4. A novel SARDINet architecture is first introduced as a relevant translator network capable of limiting the effect of SAR distortions in an urban context. This architecture is then further improved and extended to rural and mountainous regions by implementing different fusion and decoding strategies. Since the latter improvements are aimed at increasing the relevance of the reconstructed patterns but not their credibility, the chapter ends with further experiments to enhance the contrast of the reconstructed images.

Modality translation is then used for fast detection of landslide collapses in Chapter 5. After a first set of experiments based on change detection neural networks, the translator network is further trained with additional guidance. The objective of such guidance is to increase the relevance of the patterns with respect to a downstream task of interest. In particular, the TCL is introduced to allow the reconstruction of more reliable land-cover patterns and to better separate each class by using an auxiliary multi-label classifier. The resulting translations are further refined by switching the classification task to the slope instability change detection task. The goal of the so-called TCDL is to ensure that the translated image is comparable to a pre-event image and contains actual slope instability patterns to facilitate their detection.

Chapter 6 focuses on slow moving areas that are not visible on most standard remote sensing images. The geomorphological guidelines and the novel ISSLIDE dataset are first presented to better understand the target task of the chapter. Standard segmentation networks are applied to this ISSLIDE dataset to demonstrate the complexity of distinguishing between moving areas and counterfactuals, but also the non-transferability of standard pre-trained networks for such a task. The novel ECSPLAIN framework is then proposed, to account for the guidelines in the training. By combining the detection and localization of moving areas through a constraint set on the explanation of the network, the framework is able to combine explainability, lightness and efficiency. Its applicability is further demonstrated by enriching ISSLIDE with new susceptible moving areas.

A remainder of the main conclusions of the thesis as well as the gates that it opens for the future are given in a final additional Chapter 7.

Chapter 2

Geoscience and Remote Sensing Context

2.1 Introduction

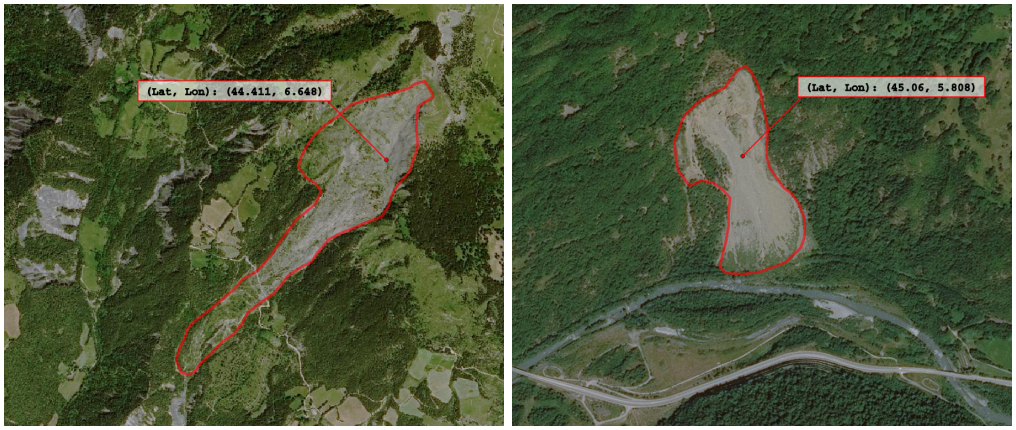
This Chapter describes the two major components of this research, slope instabilities and remote sensing images. The objective is to give the reader an understanding of the objects of study. Therefore, Section 2.2 provides a description of the three classes of phenomena under study. In addition, the Section gives an overview of the different monitoring stages used to mitigate or prevent catastrophic collapses. Standard monitoring methods based on physical parameters and *in situ* measurements are also presented in order to better understand the mechanisms and triggering factors of slope instability, and to highlight the drawbacks of such methods. The second objective is to introduce remote sensing images as a tool to overcome these drawbacks. Therefore, an accessible description of optical and Synthetic Aperture Radar (SAR) images is proposed in Section 2.3. The standard use of each modality applied to slope instability monitoring is also detailed to better understand their strengths and weaknesses. Finally, Section 2.4 introduces SAR interferometry as a way to leverage SAR images to identify small deformations.

2.2 Slope Instability Processes

2.2.1 Phenomena of Interest

2.2.1.1 Sudden landslides

Sudden landslides are of particular interest in the literature because of the catastrophic toll they can cause. Indeed, due to their very high to extremely high velocity (from 3 m/min to 5 m/s [Hungr et al., 2014]), researches mostly focus on early response to these threats. Typical responses consist of highlighting the damaged regions and providing emergency assistance to the local injured population. Most of the sudden landslide collapses are the consequences of another natural disaster event such as typhoon [Ghorbanzadeh et al., 2021], intense rainfall [Intrieri et al., 2017, Ullo et al., 2019], erosion [Tran et al., 2022] or earthquakes [Ghorbanzadeh et al., 2021]. The authors of [Varnes, 1978, McColl, 2022] list more detailed causes of landslide collapses, including natural and human factors. Overall, and based on the landslide classification available in [Varnes, 1978], its update in [Hungr et al., 2014] and [Brideau and Roberts, 2022] and the glossary in [Cruden et al., 1993], this research targets both soil slides and flow movements. The former can be described as a mass of terrain moving along a shear surface, also called surface of rupture in [Cruden et al., 1993]. On the other hand, flow movements are not homogeneous phenomena. Indeed, despite the fact that they form a continuous mass with different materials (*e.g.* earth, rocks, ice), their materials are not consolidated into a rigid block. Such behavior can be assimilated to that of viscous fluids. Depending on the coarseness and



(a) Slide and mudflow of La Valette, Alpes de Haute Provence, France.

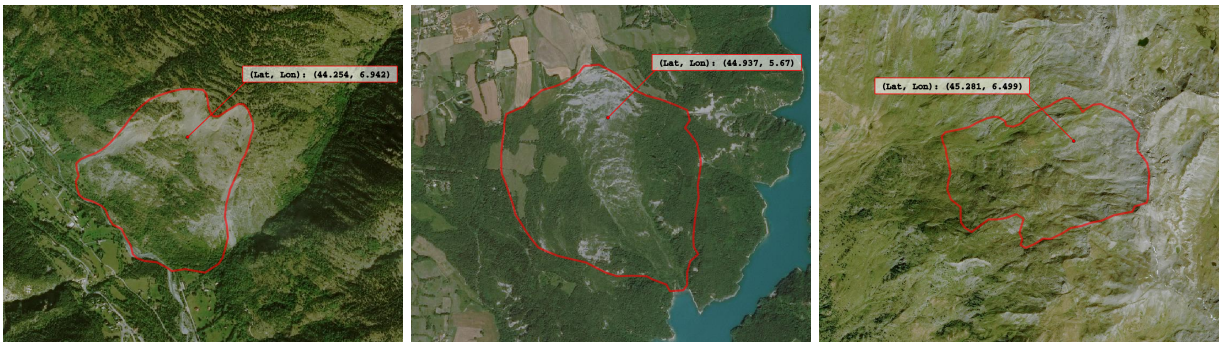
(b) Debris flow of the Ruines de la Séchiliennes, Isère, France.

Figure 2.1: SPOT6-7 images acquired in 2023 of past sudden landslides.

wetness of the materials, flows are generally classified as debris flows or mudflows. Figures 2.1a and 2.1b give example of a slide which turns into mudflow and an example of a debris flow, respectively. The La Valette landslide in Figure 2.1a could be characterized as a complex landslide based on the classification in [Varnes, 1978], as it combines both a slide on its head and a mudflow on its tongue. Although sudden landslides are too fast for real-time monitoring, precursors to their collapse can occur hours, weeks or even months before the catastrophic event [Lacroix et al., 2020]. In particular, a typical precursor is the appearance of small ground motions [Desrues et al., 2019, Intrieri et al., 2017, Lacroix et al., 2018] that make the area behave similarly to a slow-moving landslide. Furthermore, once a landslide has collapsed, it may still show signs of movement, which requires an everyday attention to avoid additional casualties due to a late triggered collapse. The two landslides taken as examples in Figure 2.1 exhibit such slow-moving landslide behavior and are therefore kept under constant observation.

2.2.1.2 Slow-moving landslides

Due to their slow movements from millimeters to tens of meters per year, slow-moving landslides can be monitored by both remote sensing images and periodic *in situ* measurements. In this way, their behavior can be better predicted and their susceptible failures can be mitigated to prevent catastrophic tolls. These



(a) Huge slide of la Clapière, Alpes Maritimes, France.

(b) Slide of les Harmalières, Isère, France.

(c) Mountain slope deformation near the Col de Montfiot, Savoie, France.

Figure 2.2: SPOT6-7 images acquired in 2023 of slow-moving landslides.

landslides are mainly characterized as slides [Varnes, 1978] and slope deformations [Hungr et al., 2014]. The latter consist of very slow rates of motion with no apparent rupture surface, only cracks and scarps on the surface. Figures 2.2a and 2.2b are taken from two slides that show a clear rupture surface while Figure 2.2c illustrates the case of a mountain slope deformation without evidence of motion at the surface, but which can be detected using the methodology developed in Section 6.2.

Slow-moving landslides require special attention, as they can show a discontinuous range of movements in time [Lacroix et al., 2020]. Indeed, due to seasonal variations, landslides can *suspend* their activity during the dry seasons while being *reactivated* during the rainy seasons. Such phenomena can thus exhibit daily velocities ranging from millimeters to tens of meters per year within a single year [Lacroix et al., 2020]. Meanwhile, a landslide that shows no activity for more than 12 months, but is still affected by potential triggering causes, is considered *dormant* [Cruden et al., 1993] because it can be awakened/reactivated at any time. On the other hand, landslides whose causes have been suppressed, either naturally or by human infrastructures, are considered *abandoned* or *stabilized*, respectively.

Most of these sliding areas are made of water-sensitive materials. As a result, their movements are particularly affected by periods of rainfall and snowmelt. On the one hand, rainfall causes ground surface erosion and fluidizes the dry underground layers that follow the natural flow of water and carry the overlying layers down the slope. On the other hand, during freezing and snowing periods, water is prone to turn into ice in the cracks of the landslides. Such transformation can increase their size and weaken the underlying materials that flow down the slope along with the water resulting from melting ice and snow.

2.2.1.3 Rock glaciers

As mentioned in [Hungr et al., 2014], all slope instabilities may not consist only of earth and rocks. In particular, ice can be a key material that causes movements due to seasonal thawing, as illustrated by the regular advance and retreat of glaciers each year. This research focuses on another phenomenon that couples ice and rock debris: rock glaciers.

Rock glaciers have been widely studied since 1905 [Martin and Whalley, 1987] for their water storage capacity, which can substitute to glaciers [Martin and Whalley, 1987, Giardino et al., 2014, Jones et al., 2018, Jones et al., 2019]. Nevertheless, with the first documented rock glacier collapse occurring in 2006 [Bodin et al., 2017], geomorphologists increasingly consider rock glaciers as potential threats that require more attention and further studies to understand their collapse mechanisms and avoid catastrophic tolls.



(a) Rock glacier of the Col du Valonnet, Savoie, France.

(b) Rock glacier of the Mont Vallon, Savoie, France.

Figure 2.3: SPOT6-7 images acquired in 2023 of rock glaciers.

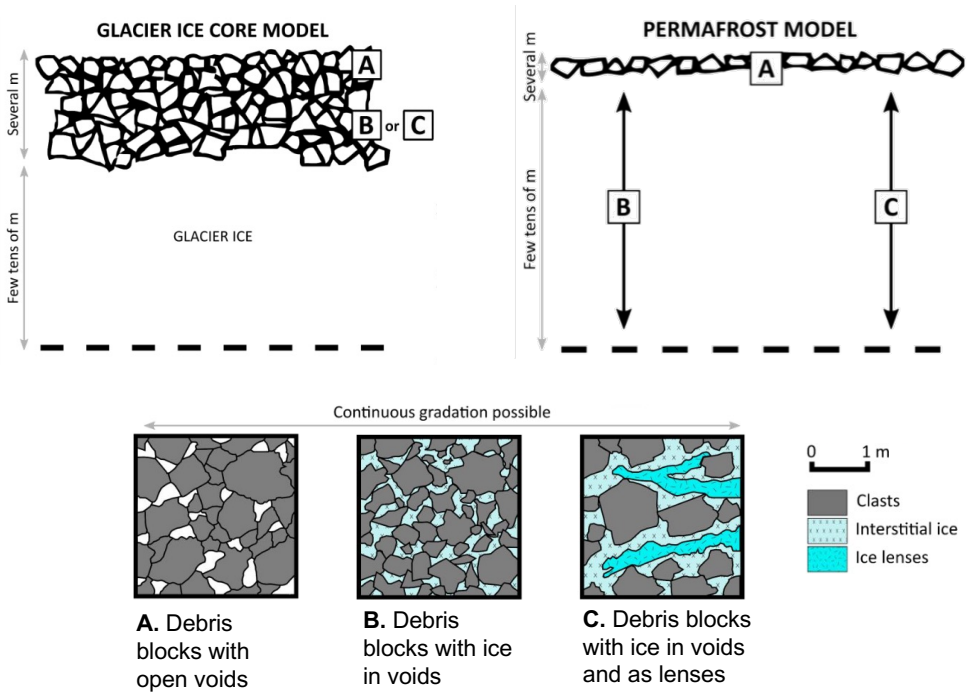


Figure 2.4: Scheme of the two models of rock glaciers: glacier ice cored and permafrost. Image modified from [Martin and Whalley, 1987] and [Jones et al., 2019]

They are easily recognized by the ridges and furrows created by the accumulation of debris on their snout. Figure 2.3 illustrates these characteristic optical patterns on two rock glaciers from the Vanoise Massif, France. Rock glaciers are not properly considered as "Landslides". Indeed, although the authors of [Hungr et al., 2014] include ice as fall or avalanche material, it is excluded from the slide category. The classification of rock glaciers, and even the term rock glacier itself, is widely debated among geomorphologists [Martin and Whalley, 1987, Giardino et al., 2014, Jones et al., 2019]. Since the goal of this thesis is to detect instabilities from a satellite perpective, we do not intend to fuel this debate. Therefore, only a brief introduction to the two main types of rock glaciers is given below and illustrated in Figure 2.4.

Glacier ice cored model: This model (top left of Figure 2.4) assumes that a rock glacier is derived from an actual glacier. After a series of debris falls onto the glacier, both ice and rocks have coalesced into a continuum [Deline et al., 2022]. Therefore, the model considers three types of layers, separating the surface from the bedrock: an upper rock-only layer, then a mixture of rock and ice, and finally a deep glacier ice-only layer.

Permafrost or ice-cemented model: This model considers rock glaciers to be independent of glaciers, as visible on the upper right corner of Figure 2.4. Their formation is mainly due to the accumulation of rock debris at high altitudes. Then, as a result of intense periods of freezing and snowing, the rock debris are cemented by the ice and form a thick layer on the bedrock. A thinner rock-only layer isolates the ice from the surface and prevents it from thawing.

As with landslides, rock glaciers can have different states of activity depending on the amount of ice remaining in the deepest layers. Indeed, if no ice remains inside the rock glacier but geomorphologic surface characteristics persist, rock glaciers are considered *relict* - which can be associated with the abandoned landslide activity state. On the other hand, *intact* rock glaciers still contain ice and are classified as *inactive* or *active*. The former show no signs of movement, which can be due either to the fact that the ice has melted (especially in summer periods) or because they are too far from a steep slope that prevents gravity from making the area slide. In contrast, *active* rock glaciers usually move centimeters to decimeters per year. The

main cause of their movement is the melting of the underlying ice caused by the rise in temperature at the surface or the flow of rain in the internal cracks. However, recent study cases show peaks in velocity of up to several meters per year. These recent accelerations may be the cause of an increase in rock glacier collapses [Bodin et al., 2017, Kofler et al., 2021, Marcer et al., 2020]. Such collapses are of particular concern during intense rainfall events as they can turn into mudflows and devastate villages down the slopes [Marcer et al., 2020].

2.2.2 Monitoring Stages

In order to avoid a huge toll due to slope failures, different monitoring steps can be set up depending on the activity state of the instabilities. Indeed, for prevention purposes, slope instability zoning can be applied to assess the locations and dangers of the phenomena. Then, if triggering conditions are predicted in the near future, early warning systems can be used to predict the most likely scenarios and take action before disasters occur. Finally, post-event landslide inventories are used specifically to assess the damage, send emergency assistance to the site and ensure its safety.

2.2.2.1 Slope instability zoning

The best way to mitigate slope instability hazards is to assess their susceptible locations and associated risks before an event occurs. According to [Fell et al., 2008], such an assessment is known as slope instability zoning. Effective zoning can be divided into three distinct tasks.

- ***Susceptibility Zoning***: The first step is to locate susceptible slope instabilities. Based on geomorphology, topography, climatic conditions, water-based indices or distance from human infrastructure, the aim is to produce maps that highlight the locations that are susceptible to host slope instability. This step makes it possible to identify the causes of instability in an area and to install infrastructures to avoid the danger.
- ***Hazard Zoning***: Hazard zoning consists of evaluating the probability of collapse of the slope instabilities identified in the susceptibility map. Several factors can be examined, such as the presence of nearby instabilities, changes in activity state, previously known events or the estimated volume and velocity of the phenomenon. This map provides a better understanding of the potential occurrence of slope failures per year and is a key tool to help authorities make decisions to protect the most vulnerable areas.
- ***Risk Zoning***: The measurement of risk is not directly related to the phenomenon but to the damage it can cause in terms of human lives and destruction of infrastructures. Therefore, the coupling of the previous hazard zoning map with an external land use map allows to identify the villages, roads or dams at risk and to prepare their evacuation or closure in case of an increase in hazard.

Based on these different maps, local authorities can make appropriate and conscientious decisions to prevent the damage caused by disasters before they happen.

2.2.2.2 Slope instability early warning systems

Despite the creation of slope instability zoning maps, some areas cannot be stabilized with human resources because of poor accessibility that could endanger workers, because of the proximity to existing infrastructures, or even because of the lack of financial resources. As a result, many efforts have been made to develop early warning systems to predict collapses and evacuate the local population before a failure occurs. The vast majority of early warning systems consists of sending alerts when pre-defined thresholds are exceeded. Although rainfall thresholds are widely used in the field, several other factors can be measured and monitored to provide an early warning of slope failures. In particular, the authors of [Intrieri et al., 2013] propose a general instrumentation flowchart depending on the type of slope instability being targeted, its internal characteristics, the potential causes of its movement and the available budget. The aim is to help the

authorities to install the most reliable system according to the situation. Nevertheless, as stated in [Guzzetti et al., 2020], too few early warning systems are implemented worldwide. Therefore, the authors suggest 30 different recommendations to improve the development, dissemination and performance of early warning systems.

2.2.2.3 Slope instability inventory

Due to a lack of zoning or early warning systems, slope instabilities continue to have devastating effects around the world every year. Slope instability inventories are essential for early response and emergency assistance after an event. An inventory can refer to several definitions in the field, as detailed in [Fell et al., 2008]. Three levels of inventories are considered and can be summarized as:

- Basic inventories, with the coordinates, class, volume and date of the disaster.
- Intermediate inventories, in which a more detailed mapping is carried out by delineating the boundaries of the phenomenon and identifying its recent activity states.
- Sophisticated inventories, which consist of a deeper historical context and advanced geomorphological analyses.

Inventories are mostly applied *a posteriori* of an event to quantify the damage caused by the disaster with bitemporal change detection approaches [Nagendra et al., 2022, Nava et al., 2021] or with pattern recognition algorithms applied to a single post-event image [Ghorbanzadeh et al., 2019, Ghorbanzadeh et al., 2021, Yang et al., 2022a]. Yet, with an appropriate methodology, inventories can be constructed prior to any collapse by targeting the localization of slow-moving slope instabilities as in [Bertone et al., 2022, Jones et al., 2018, Rouyet et al., 2021]. Such inventories can thus be used as input for instability zoning to better predict potential threats to populations. Although these inventories help provide early warning or early response to slope failures, they are also critical for researchers. Not only will this mass of information be used *a posteriori* to better understand the internal mechanisms and kinematics of slope instabilities, but with the exponential development of machine learning methods, it can also be used to strengthen the reliability and accuracy of automatic predictions.

2.2.3 Standard Approaches

Despite the fact that slope instability phenomena have been studied for a long time [Varnes, 1978, Martin and Whalley, 1987], the subsurface mechanisms and kinematics are still a matter of concern. In particular, with the great development of measurement tools, researchers are focusing their interest on the search for evidence that can help predict future collapse. To this end, numerous approaches based on physics and triggering factors are being implemented to better interpret, understand and predict signs of the presence of slope instability and/or susceptibility to collapse.

2.2.3.1 Physical models

Physical models are mainly used to predict and determine the collapse date of slope instabilities based on internal and external conditions. All of them are based on the study of the behavior of the particles belonging to the phenomenon, assuming that the latter are large enough in comparison with the particles, this hypothesis is also known as the *infinite slope model*. Under this condition, the authors of [Liao et al., 2011] propose the Transient Rainfall Infiltration and Grid-based Regional Slope-stability model (TRIGRS). It consists of generating the factor of safety of a given area based on topography, soil hydraulic diffusivity and rainfall duration and intensity. A factor of safety less than 1 is then associated with slope failure. In [Michel et al., 2014], the authors propose to compare two models derived from the infinite slope model and the steady-state hydrological model - which estimates soil moisture. On the one hand, they compute the SHAllow Landsliding STABility model (SHALSTAB), which rewrites the steady-state hydrological model by integrating the infinite slope hypothesis. In this way, they define two extreme conditions that determine

whether a slope is unconditionally stable or unstable. If neither is true, then the slope is susceptible to failure. On the other hand, the Stability INdex MAPping (SINMAP) combines both the SHALSTAB model and the topographic characteristics of the area to generate a stability index that indicates the probability that a given location is stable. Similarly, the authors of [He et al., 2016a] combined the Coupled Routing and Excess Storage model and the Slope-Infiltration-Distributed Equilibrium model into CRESLIDE. It leverages the powerful capabilities of the first model to estimate effective rainfall infiltration and integrates it with the second model to calculate a more reliable factor of safety. From a spatio-temporal perspective, the Time Variant Slope Stability (TiVaSS) model is developed in [An et al., 2016]. The model not only incorporates the time-varying state into the infinite slope model but also combines it with the 3D subsurface flow model to assess slope stability in real time during an intense rainfall event.

This brief review of physical models aims to better understand the actual subsurface behaviors at stake in slope failure detection, as well as their dependence on internal and external conditions. With the growing interest and power of machine and deep learning approaches, such physical knowledge could be leveraged to select the physical parameters used as input to the models [Inan and Rahman, 2023], to evaluate their reliability, to condition the model behavior through Physically Informed Neural Networks (PINN) [Raissi et al., 2019], or to pre-process the data [Wei et al., 2021].

2.2.3.2 Machine learning on conditioning factors

The selection of accurate conditioning factors for slope instability predictions is a key step for accurate susceptibility mapping [Fang et al., 2020, Wei et al., 2021, Inan and Rahman, 2023], hazard mapping [Aslam et al., 2021] and early warning systems [Guzzetti et al., 2020]. Since the latter aim at detecting imminent collapses, they are mainly based on triggering factors such as rainfall intensity, seismicity, climate and geology according to the 26 systems reviewed in [Guzzetti et al., 2020]. On the other hand, slope instability zoning approaches rely mainly on topography (*e.g.* slope, curvature, elevation, aspect), geology (*e.g.* soil type, lithology), land cover (*e.g.* vegetation indices, distance to roads and rivers) and hydrology (*e.g.* water indices, stream power, wetness index). Sixteen indices are selected in [Aslam et al., 2021] to be input to a Convolutional Neural Network (CNN) for feature extraction. These features are then classified by four different machine learning approaches to generate hazard maps in Northern Pakistan. Following the same concept, the authors of [Fang et al., 2020] use 16 factors to generate susceptibility maps in the territory of Yongxin in China. To take advantage of the physical explanations provided by the TRIGRS model, the authors of [Wei et al., 2021] compare the performance of three CNNs. The first is trained on twelve conditioning factors, similar to the methods mentioned above. The second replaces the lithology with the factor of safety resulting from the TRIGRS model. The third network is trained only on a subset of samples preselected according to the factor of safety from TRIGRS. Indeed, areas with a high factor of safety are by definition stable and thus useless for the training stage. This final CNN shows outstanding results compared to its counterparts, illustrating that the fusion of physics and deep learning can take the best of both worlds. Another physics-friendly approach is to explain the results of neural networks. In particular, the authors of [Inan and Rahman, 2023] implement a susceptibility mapping approach based on an explanation method used to highlight the factors that are actually relevant among fifteen inputs during the decision process of five machine learning classifiers.

Therefore, the integration of physics into machine learning approaches for slope instability monitoring definitely shows encouraging results. It not only exploits the powerful generalization capabilities of machine learning methods, but also makes them understandable and consistent with physical models.

2.3 Remote Sensing Images

The main issue in requiring triggering factors for instability detection is their availability. Indeed, depending on the country, the region or even the specific location of the phenomenon, some factors may be missing. In particular, precipitation records are critical for standard physics-based models, as well as the subsurface geology. In addition, these factors should be regularly updated to ensure efficient and reliable predictions under changing conditions. Finally, these approaches are mainly applied locally, while a given instability is

susceptible to generate or be generated by another instability. Such cascading events require a more global view to predict any susceptible threats. To overcome these problems, remote sensing images are widely exploited [Zhao and Lu, 2018, Lissak et al., 2020, Ray et al., 2020, Mondini et al., 2021, Lacroix et al., 2022]. This study focuses on optical and radar images acquired by Sentinel-1 and Sentinel-2. A detailed description of each modality, their standard use for slope instability monitoring and their respective drawbacks is given below.

2.3.1 Optical Images

2.3.1.1 Multispectral images

In the field of satellite remote sensing, the multispectral sensors used to obtain optical images are passive sensors: they capture the spectral characteristics of a given object by sensing the signal emitted by an external source and reflected by the object. The sensors are designed to discriminate between different spectral bands, and thus generate a multichannel image. If standard photographs focus mainly on detecting wavelengths in the visible range (400 to 700 nm) to generate Red Green Blue (RGB) channels, multispectral sensors can detect a multitude of channels (from 10 to 30) and can reach up to hundreds of channels with specific *hyperspectral sensors*.

Among the many multispectral satellites orbiting around the Earth, different types of image can be acquired depending on the spatial resolution and the number of spectral bands targeted. In this thesis, spatial resolution is considered as *coarse* when the pixel size is greater than 30m, typically, the MODIS satellite acquires 36 spectral bands every 2 days with a resolution from 250 to 1,000m. Then *medium* resolution ranges from 5 to 30 meters such as the 11 bands of Landsat-8 at 30-meter spatial resolution in the visible domain acquired every 16 days, the Sentinel-2 images detailed below or the SPOT6-7 satellite which provide 6-meter resolution RGB and Near- InfraRed (NIR) channels every 26 days. The latter are close to *high* resolution images which resolution ranges between 1 and 5 meters. For this reason, the *Institut National de l'information Géographique et forestière* (IGN) provides an annual concatenation of SPOT6-7 image tiles covering France, mainland[1]. Another example of high resolution satellite orbiting over France is the Pléiades constellation which acquires RGB-NIR images at 3-meter resolution. The rare satellites capable to reach below 1-meter resolution are finally considered as *very-high* resolution satellites such as the Maxar's WorldView-2 satellite which acquires panchromatic and multispectral images at 0.46 and 1.85 meter-resolution, respectively. Another constellation ensures the acquisition of a daily image with a variable resolution. The Planet constellation consists of three types of satellite: Planetscope with a resolution of 3 meters, RapidEye at 5 meters and Skysat at 0.5 meters, which can all acquire RGB-NIR images. This thesis focuses on the Sentinel-2 constellation launched by the European Space Agency (ESA), which consists of two satellites in orbital phase opposition: Sentinel-2A and Sentinel-2B. Each satellite has a 10-day equatorial revisit period, resulting in the acquisition of a multispectral image every 5 days. The Sentinel-2 optical sensors are calibrated to acquire 13 different bands.

- Four bands at 10m resolution to collect RGB-NIR channels;
- Six bands at 20m resolution with four in the NIR range for vegetation analysis and two in the shortwave infrared range for vegetation, soil moisture and snow applications;
- Three bands at 60m resolution for aerosol, water vapor and cirrus cloud detection.

The Level-2A Sentinel-2 products provided by ESA result from a series of post-acquisition processing steps to georeference the image, and from the application of the Sen2Cor atmospheric correction processor[2]. They consist of tiles of 110×110km with a pixel resolution of 10×10m encoded on 12 bits for the RGB-NIR bands of interest in this thesis.

[1] https://geoservices.ign.fr/documentation/services/utilisation-sig/tutoriel-qgis/wms-wmts, last visited on June 25, 2024.
[2] https://step.esa.int/main/snap-supported-plugins/sen2cor/, last visited on March 17, 2024

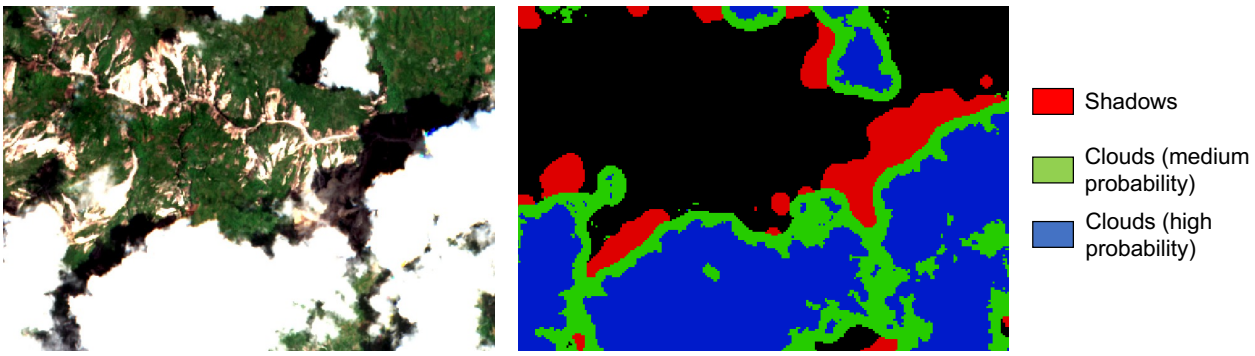


Figure 2.5: Impact of the cloud coverage on the visibility of landslides in Haiti. Image acquired by Sentinel-2 satellite on August 14, 2021.

2.3.1.2 Optical images for slope instabilities

Monitoring slope instabilities with optical satellite images is a challenging task due to their resolution. Indeed, optical correlation techniques can be applied at any resolution but are unable to detect or to monitor small instabilities on coarse resolution images [Lacroix et al., 2022]. Therefore, high-resolution images or multi-resolution approaches such as Cosi-Corr or Mic-Mac are preferred to assess the kinematics and states of activity of slope instabilities [Lacroix et al., 2022, Ray et al., 2020, Zhao and Lu, 2018]. Nevertheless, some approaches still consider medium resolution images such as Sentinel-2 (10-meter resolution) to detect precursory movements of the Harmalières landslide visible in Figure 2.2b [Lacroix et al., 2018]. The Cosi-Corr algorithm is exploited to generate displacement maps between images acquired at intervals ranging from 3 to 53 days. By their side, precursory movements are studied mainly by using high to very-high resolution images. The authors of [Delacourt et al., 2004] study QuickBird images (0.6 meter resolution) and aerial photographs to measure the velocity field of the Clapière landslide (see Figure 2.2a) using correlation techniques. Similarly, in [Desrues et al., 2019], Pléiades panchromatic images (0.7 meter resolution) are correlated to detect early failures of the Tunnel du Chambon landslide.

However, optical images are particularly useful for creating slope instability inventories. Indeed, most of the 75 rock glacier inventories from the literature and identified in [Jones et al., 2018] were annotated based on aerial or satellite optical images. In addition, the authors of [Lissak et al., 2020] review methods that use remote sensing images to assess slope instability and find that 33 approaches produce inventories using optical images acquired from 12 different satellites. Indeed, even if the resolution is still challenging for the detection of small phenomena, the inventory task focuses on radiometric sudden changes that can be easily identified despite a coarse resolution. Conversely, monitoring tasks focus on image correlations to find centimeter-scale movements which by definition are affected by the resolution of the image. This is the reason why optical images are significantly more used for inventory purpose than for slow motion detections [Zhao and Lu, 2018, Lissak et al., 2020, Ray et al., 2020].

2.3.1.3 Radiometrical issues

Another reason why optical images are underutilized for instability monitoring is their radiometric sensitivity. Due to their passive nature, satellite optical sensors rely on the sun to emit a source signal. As a result, not only are the images not available at night but they are also greatly affected by seasonal variations in solar exposure. Such variations produce a darker image with larger shadows in winter compared to summer.

The most critical issue is their meteorological sensitivity. Indeed, the presence of clouds over an area prevents multispectral sensors from imaging it. To inform users about the presence of clouds in the image, Level-2A products are delivered with a set of masks calculated by applying a series of thresholds to the different bands during the Scene Classification of the Sen2Cor process. It results in four cloud-related masks: a low, medium and high probability mask, and a cloud shadow mask. Indeed, due to the orientation



(a) Sentinel-2 image October 29, 2020

(b) Sentinel-2 image February 26, 2021

(c) Sentinel-1 image December 30, 2020

Figure 2.6: Study case of the Gjerdrüm landslide nearby Ask, Norway (images from EO browser[4]). The red line shows the extent of the landslide.

of the sun, the shadows coming from the clouds can be visible on the images and hide the regions of interest such as sudden landslides in Haiti, as visible in Figure 2.5.

Not only can cloud cover partially obscure the image and prevent the localization of landslides, but it can also make the images useless. Indeed, as mentioned in Section 2.2, slope instability failures or reactivations are often triggered by intense rainfall. Such a meteorological event implies the presence of a large cloud cover over the area, making the entire multispectral acquisition useless. As a matter of fact, Figure 2.6 illustrates the case study of Gjerdrüm[3] [Tran et al., 2022] where a huge landslide destroyed the south of the city of Ask, Norway. Figure 2.6b shows the first optical image available after the disaster. Compared to the pre-disaster image in Figure 2.6a, the landslide is clearly visible with significant damage to forests, rivers, and some neighborhoods of Ask. However, the post-disaster optical image is taken on the 26th of February 2021, while the actual landslide occurred two months earlier, on the 30th of December 2020. Such a delay, caused by extensive cloud cover, prevents early damage assessment and limits the organization of emergency assistance by the authorities. The comparison of both optical images is even more challenging since snowfall occurred between the acquisitions. Therefore, a straightforward change detection approach would consider the entire image as changed while the focus should be only set on the landslide location.

2.3.2 Synthetic Aperture Radar Images

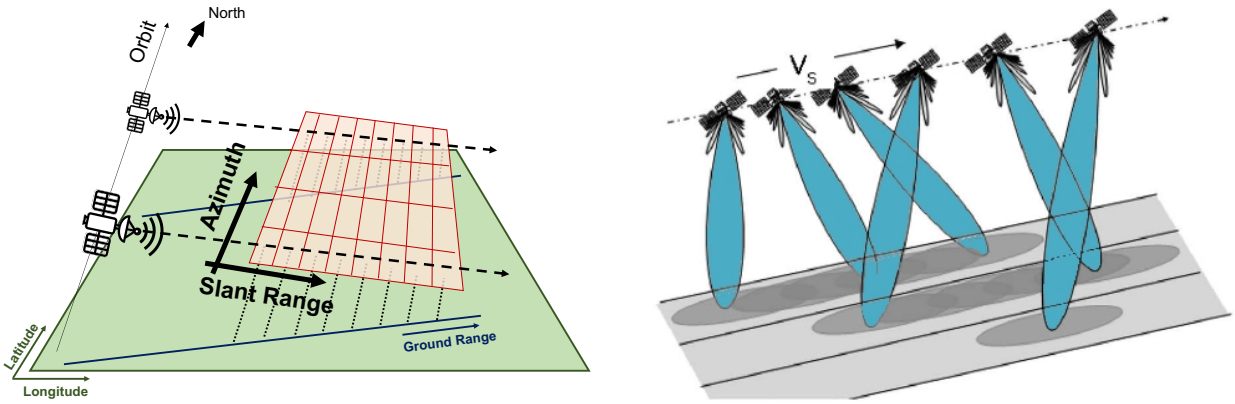
The case study of the Gjerdrüm landslide is also a perfect example to illustrate the relevance of SAR images in detecting slope instabilities. Indeed, on the SAR image in Figure 2.6c which was acquired immediately after the landslide failure, the bright pixels correspond exactly to the location of the landslide, providing critical information for emergency response. This case study is not isolated: a large number of landslides caused by intense rainfall cannot be detected early with optical images due to cloud cover. Therefore, radar images are exploited for early assessment of the extent of the event and for early response to the disaster. The characteristics, strengths and weaknesses of SAR imagery are discussed below.

2.3.2.1 Radar images

Unlike their optical counterparts, satellite Synthetic Aperture Radar (SAR) sensors are active. Specifically, an electromagnetic wave is first sent from the sensor to the target. The backscattered signal $z = Ae^{i\varphi}$ is then detected by the sensor which records both its amplitude A, corresponding to the amount of backscattered signal, and its phase φ, related to the delay between the emission and reception modulo 2π. The complex image resulting from the SAR synthesis - also named Single Look Complex (SLC) product - can be used to exploit the phase difference between two acquisitions by interferometry (see Section 2.4.2). Another

[3]https://www.regjeringen.no/contentassets/3dadc8f7fad94608861163fa524023c0/no/pdfs/arsakene-til-kvikkleireskredet-i-gjerdrum-2020.pdf, report of the disaster (Norwegian version only), last visited on March 17, 2024.

[4]https://apps.sentinel-hub.com/eo-browser/, last visited on March 17, 2024.



(a) Radar geometry image acquisition for an ascending pass.

(b) TOPSAR illustration from the ESA[5].

Figure 2.7: Sentinel-1 image acquisition process.

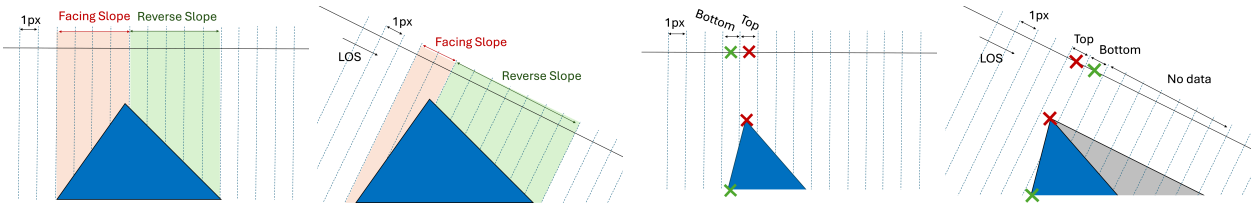
processing is applied by the ESA to produce an amplitude-only Ground Range Detected (GRD) product as illustrated in Figure 2.6c. Depending on the configuration of the SAR sensor, the images may reflect different characteristics of the same underlying area.

Polarization: The available polarization of the emitter and the receiver can be used to acquire up to four different images: two co-polarized bands HH and VV and two cross-polarized bands HV and VH. In the co-polarization mode, the emitter and receiver have the same polarization (Horizontal: H or Vertical: V), while in the cross-polarization mode they are different. Therefore, an HV image is obtained by emitting a Horizontally polarized wave and receiving the Vertical component of the backscattered signal. Full-polarimetric sensors are designed to acquire all bands while most sensors are only dual-polarimetric and acquire VV+VH or HH+HV images.

Wavelength: The wavelength and the available bandwidth used for acquisition determines not only the resolution of the image, but also the visibility of certain structures in the image. In particular, low frequency bands such as L-band (1-2 GHz) and S-band (2-4 GHz) are less sensitive to small objects (*e.g.* the leaves on the trees) but have a coarser resolution than high frequency bands such as X-band (8-12 GHz). Therefore, depending on the application, one can use L-band SAR images to detect objects under the canopy or X-band images to monitor the deforestation process.

Several satellites and constellations of satellites have been launched over the last 24 years with different bandwidths: L-band with ALOS-1 (2006) and -2 (2014), C-band with ENVISAT (2002) and Radarsat (2007), or X-band with TerraSAR-X (2007) and TanDEM-X (2010). The future NISAR planned to be launched in 2024 is even supposed to acquire both L-band and S-band. But the most widely used SAR constellation is Sentinel-1. Indeed, before 2021, the constellation consisted of two identical dual-polarization C-band satellites, Sentinel-1A and Sentinel-1B, launched in 2014 and 2016, respectively, and traveling in orbital phase opposition. The two satellites covered almost the entire Earth (excluding the South Pole) with a 12-day repetitivity in the same configuration. This routine allowed the acquisition of a new image every 6 days over Europe due to their orbital phase opposition. In addition, the Sentinel-1 satellites could acquire images during both ascending (from the South Pole to the North Pole) and descending passes (from North to South). Thus, using all available images, the same European area can be seen even more frequently which is a significant advantage over other SAR missions. However, in December 2021, Sentinel-1B experienced issues that have made it unavailable for image acquisition, leaving Sentinel-1A alone for SAR acquisition. Sentinel-1C is scheduled for launch in 2024 to compensate and replace the defective satellite.

[5]https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-1-sar/acquisition-modes/interferometric-wide-swath, last visited on March 17, 2024.



(a) Ground geometry - compression, dilation

(b) Radar Geometry - compression, dilation

(c) Ground geometry - Layover, shadows

(d) Radar geometry - Layover, shadows

Figure 2.8: Comparison of a steep slope area in ground and radar geometry. Geometrical distortions appear on the latter.

Contrary to standard optical imaging, SAR images are acquired in **radar geometry** rather than the typical **ground geometry**. This means that a 1-meter resolution pixel in a radar image does not correspond to an actual 1-meter distance at the surface of the Earth. Indeed, SAR images are sampled in slant range along the Line Of Sight (LOS) of the satellite, as depicted in Figure 2.7a. Depending on the topography of the underlying terrain, some areas may be compressed or dilated, causing geometrical distortions in the acquired image (see the specific paragraphs about geometrical distortions below for more details). The second dimension of the image - called azimuth in radar geometry - is sampled in distance along the orbital path of the satellite.

The Sentinel-1 sensors are designed to image the Earth in four different acquisition modes.

- Interferometric Wide (IW) swath mode: This configuration is widely used for imaging continental areas. It uses the Terrain Observation with Progressive Scans SAR (TOPSAR) method to compute 3 sub-bands of images called swaths. As depicted in Figure 2.7b, it consists in acquiring the three swaths one after the other while the satellite keeps traveling, ensuring an azimuth overlap between the acquisitions for continuity. Each swath is acquired by sending a *beam* of electromagnetic waves to the ground for a given azimuth interval resulting in a *burst* of data which are then concatenated to form the swath. The spatial coverage of the image is 250 km wide with a pixel spacing of 2.3×14.1 meters.

- Extra Wide swath mode: The same TOPSAR technique is used to compute 5 swaths covering 410 km. This approach trades resolution (7.9×43 meters in range-azimuth) for wider spatial coverage to monitor oil spills and ice in specific ocean areas and in polar zones.

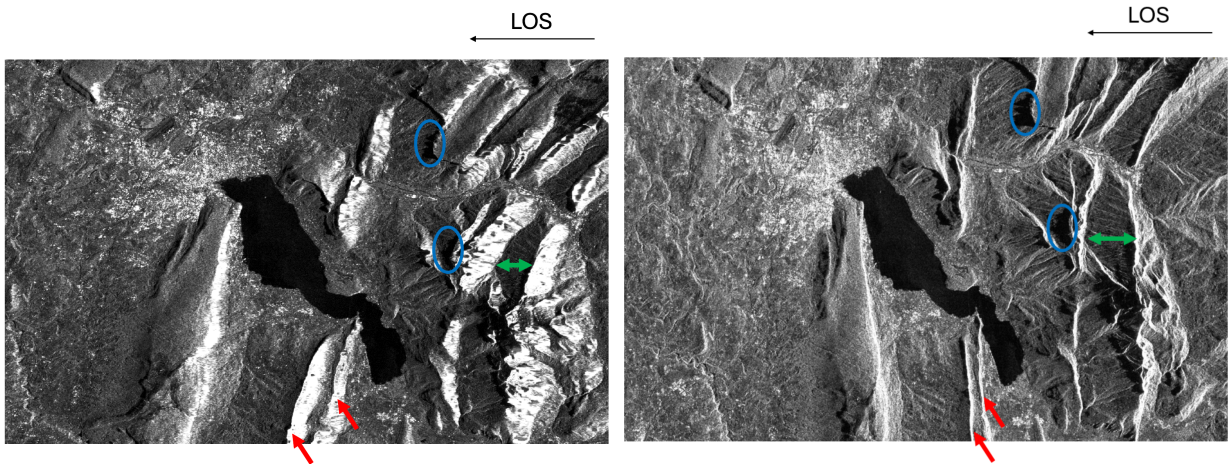
- Wave mode: This is used specifically to map the entire open ocean by generating a mosaic of 20 km wide images with a resolution of 2×4.8 meters in range-azimuth. The acquisition is based on a "leap frog" strategy, *i.e.* the images are not acquired continuously. An image is taken every 100 km to reduce the amount of data to be stored while covering the ocean in regular spatial steps.

- Stripmap mode: This configuration acquires a single more precise swath with a resolution of 1.7×4.3 meters in range-azimuth but is only 80 km wide. Its use is specific to some islands, but it can also be involved in case of extreme emergency to support actions after major disasters.

As mentioned in Section 2.2, this work focuses on continental slope instabilities, which mostly occur in mountainous regions. Therefore, the IW mode acquired with the TOPSAR technique is used for the data considered in this thesis.

2.3.2.2 Geometrical distortions

One of the major challenges in detecting slope instability on SAR images is the location of such phenomena. Indeed, movement is likely to occur in areas with slopes greater than 30°. However, these steep slopes cause distortions in the radar products [Kaushik et al., 2021] due to their particular acquisition geometry (see



(a) Orthorectified SAR image, *i.e.* in ground geometry (b) Original GRD image from the ESA, *i.e.* in radar geometry

Figure 2.9: SAR distortions in mountainous areas nearby Annecy. Image from Sentinel-1 on March 13, 2023.

Figure 2.7a). Four distortions are of particular interest in this work.

Foreshortenings and Dilations: These two distortions illustrated in Figure 2.8b are the most common distortions in mountainous areas. The former are caused by slopes facing the radar sensor, while the latter occur on the backside slope of the mountain. As depicted by the red highlighted area in Figures 2.8a and 2.8b, the facing slope is only imaged within two pixels in the radar geometry, while it should extend over four pixels in the ground geometry, in the schematic example. This compression/foreshortening issue results in a critical loss of data because the value of a SAR pixel results from the sum of the backscattered signals from a larger area than standard flat regions. Therefore, the facing slope radar pixels have a lower resolution than other pixels in the SAR image when projected onto the ground geometry. Conversely, the dilations oversample opposite slopes, as illustrated by the green areas in Figures 2.8a and 2.8b. Thus, these radar pixels have a higher resolution than other pixels in the SAR image when projected onto the ground geometry. The red and green arrows in Figure 2.9 indicate some slopes affected by foreshortenings and dilations, respectively, around Annecy.

Layovers: Foreshortenings are particularly created by gentle slopes, but if the slope exceeds the orthogonality with respect to the LOS of the satellite, the foreshortening effect is transformed into a layover. Indeed, in such a configuration, due to the height of the imaged 3D objects (*e.g.* mountains for coarse resolution or buildings for high resolution), the emitted electromagnetic wave reaches the top of the object before its base. Therefore, the first backscattered signal captured by the sensor is the top of the object which is consequently imaged before its base. Figure 2.8d illustrates the impact of layover on a SAR image compared to a standard ground geometry sampling in Figure 2.8c. This distortion creates strong unrecoverable artifacts in the SAR image that can mislead downstream decisions. Figure 2.10 shows the extent of this phenomenon in an urban context. All the bright circles in the radar image actually belong to the upper tanks in the optical image, as illustrated by the red arrows. Thus, a critical shift of tank locations is observed in the radar modality.

Radar Shadows: The shadows result from the lack of backscattered signals in certain areas. Due to the presence of steep slopes or huge 3D objects, the propagation of electromagnetic waves can be interrupted by such obstacles, preventing them from reaching the areas behind, as schemed in Figure 2.8d. These hidden regions are therefore considered by the sensor as areas with no backscattering, and thus with very low values in the generated radar image. Examples of this phenomenon are circled in blue in Figure 2.9. Shadows are particularly troublesome because no signal is collected over these areas. In addition, they can be easily confused with water bodies such as lakes and rivers which are highly reflective in the specular direction. Indeed, as illustrated in Figure 2.9, the lake of Annecy does not backscatter any signal towards the sensor and appears with a very low amplitude in the radar image. Such confusion can be the source of many errors

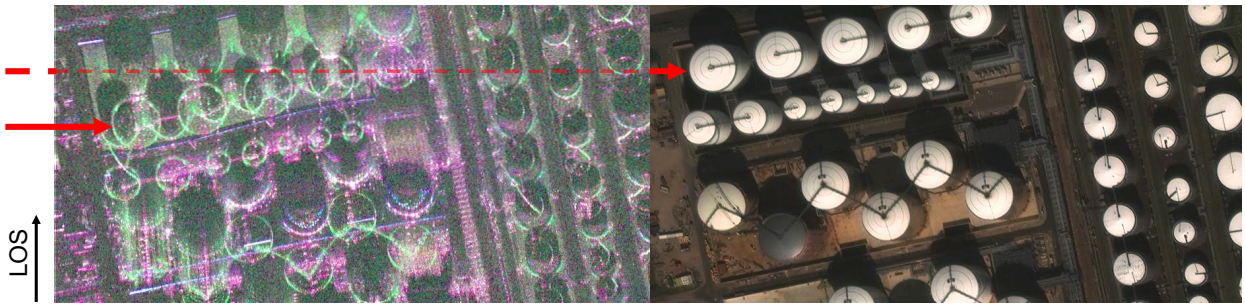


Figure 2.10: SAR distortions in urban areas in Rotterdam, the Netherlands (sample image from SpaceNet6 [Shermeyer et al., 2020]). SAR image is RGB colored as [VV, $\frac{VH+HV}{2}$, HH].

as experienced in Section 4.4.1.

Although the lack of data due to shadows cannot be artificially restored and the mixture of backscattered signals due to layovers cannot be unmixed, the irregular sampling issue caused by foreshortenings and dilations can be compensated. Indeed, by using a Digital Elevation Model (DEM) of the scanned area, SAR images can be *orthorectified* to resample them from radar geometry to ground geometry. Orthorectification uses the DEM and the incidence angle formed between the LOS of the satellite and an ellipsoid estimate of the Earth to register the radar image appropriately on the ground. This process has the advantage of resampling the radar image so that the pixel sampling is constant on the ground. The major drawback lies in foreshortened and layover areas. Indeed, since missing values can neither be guessed nor acquired, the interpolation used to upsample these pixels will propagate the mixture of signals throughout the pixels of the facing slope, producing large bright bands as indicated by the red arrows in Figure 2.9a. More critically, these distortions affect all post-acquisition processing steps, including image registration, image translation, and interferogram generation.

2.4 Radar Interferometry

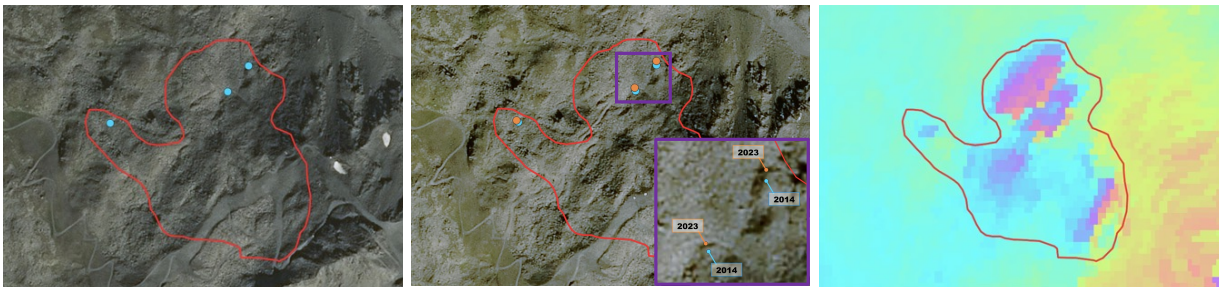
Radar interferograms are used in practice to monitor slow moving phenomena, whether landslides [Fruneau et al., 1996], volcanic deformations [Anantrasirichai et al., 2018] or rock glaciers [Bodin et al., 2018]. This Section aims to provide a better understanding of the motivations, the generation process and the standard applications of the SAR Interferometry (InSAR).

2.4.1 Exploring SAR Phase

The main idea behind SAR interferometry is to exploit the phase of the radar images. Because of its randomness resulting from the coherent sum of the responses of the distributed scatterers, its potential can only be fully exploited by combining it with another SAR image to generate phase difference images called interferograms as illustrated in Figure 2.11c.

Indeed, from a physical point of view, the phase of a radar image is directly related to the delay between the emission of the signal by the satellite and the reception of the backscattered signal. Therefore, under perfect and unchanged conditions, two successive SAR images should have identical phases while a ground motion towards or away from the satellite will shorten or extend the delay between the emission and reception, respectively. The displacement Δd between the two passes can then be characterized from the observed phase difference $\Delta\varphi$ and the wavelength λ of the satellite in Equation (2.1).

$$2 \times \Delta d = \frac{\lambda \times \Delta\varphi}{2\pi} \tag{2.1}$$



(a) SPOT 6 image - 2014

(b) SPOT 6/7 image - 2023

(c) Sentinel-1 18-day interferogram - July 02, 2018 to July 20, 2018

Figure 2.11: Illustration of the detection of small displacement rates on optical images acquired over nine years (left and center) and on an interferogram (right) computed from SAR images acquired over 18 days. The blue and orange dots are located on the same pieces of land in the optical images that moved by 7.18m, 7.7m and 9.9m from left to right between 2014 and 2023.

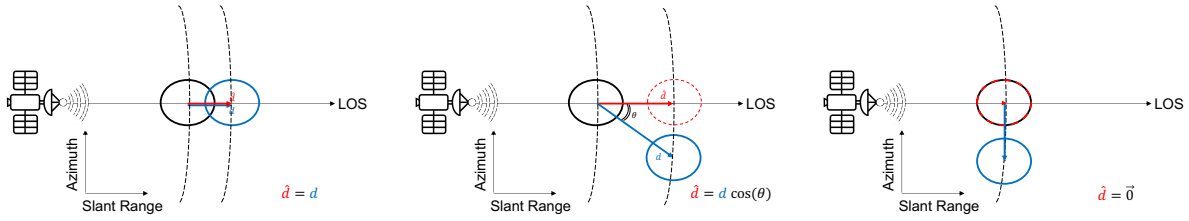
The factor 2 in Equation (2.1) reflects the fact that the phase difference is affected by both the delay of the wave as it travels from the satellite to the target and the delay as the wave returns to the satellite. Thus, the additional phase difference introduced by the actual move is accounted for twice in the observed phase difference. Considering the Sentinel-1 constellation as an example, since the emitted wavelength is $\lambda = 5.55$ cm, a maximum deformation of 2.275 cm can be observed when the phase difference of the two images is 2π. Therefore, despite the coarse resolution of the Sentinel-1 images (2.3×14.1 meters in range-azimuth on the ground), displacements of the order of centimeters can be observed thanks to InSAR products. This property explains the growing interest of geoscientists in the use of SAR interferometry. Indeed, as visible in Figure 2.11, two high-resolution SPOT 6/7 optical images were acquired in 2014 and 2023 over an active rock glacier. The three cyan and orange samples illustrate the displacement observed over the nine-year period. In particular, from left to right, the movements reach 7.18 m, 7.7 m and 9.9 m, which is smaller than the standard 10-meter resolution of Sentinel-2 images. On Figure 2.11c, an interferogram computed from two 10-meter resolution SAR Sentinel-1 images acquired within 18 days over the same area is displayed. The interferogram helps to visually identify significant changes within the annotated region (red delineation) compared to the surrounding area. This temporal baseline is chosen in Figure 2.11c for better visualization, yet this pattern can also be identified on interferograms acquired with a temporal baseline of 6 days. Therefore, the displacement that was poorly visible between two optical images acquired in 9 years can be observed in the interferogram of two SAR images acquired in 6 days. This is not only useful for accurately detecting, monitoring and estimating the displacement rate of moving areas but also for early response to susceptible threats such as sudden acceleration of the phenomena.

Nevertheless, three main issues should be considered when interpreting InSAR images: the projection of the movement, the coherence of the area and the external phase disturbances.

Projection on the Line Of Sight (LOS): Due to the peculiarity of SAR acquisitions (see Figure 2.7a), the observed displacement is not the actual ground displacement but its projection on the LOS of the satellite. Therefore, the interpretation is limited to the direction of the LOS. Indeed, as visible in Figure 2.12a, the actual displacement is only accurately estimated if it is in the same direction as the LOS. Conversely, in Figure 2.12b, only its projection onto the LOS can be measured so that a motion orthogonal to the LOS as in Figure 2.12c is invisible from the satellite perspective and cannot be estimated.

External phase differences: As mentioned earlier, the formula proposed in Equation (2.1) is reliable under perfect conditions. Nevertheless, in practice, the phase difference is affected by several external factors such as:

- The perpendicular baseline between two passes. It characterizes the difference between the two orbits taken by the satellite for each acquisition. Since the latter is not zero, the underlying terrain is not



(a) In the direction of the LOS (b) Projected on the LOS (c) Orthogonal to the LOS

Figure 2.12: Top view of movements detected by the phase difference depending on its direction with respect to the LOS. The red ($\hat{d}$) and blue (d) arrows refers to the measured and actual displacement vectors.

seen under the same eye. Therefore, this perpendicular baseline generates fringes patterns depending on the topography of the terrain, also known as topographic fringes.

- The atmospheric conditions such as the pressure, the temperature and the humidity of the air. Such variations from one acquisition to another modifies the medium of propagation of the waves, making them faster or slower. This velocity change generates an additional phase difference, also known as atmospheric fringes.

Although the topographic fringes can be mitigated with an accurate Digital Elevation Model (DEM) and the precise orbit registered by ESA, the atmospheric effect is more difficult to model in practice as it can vary significantly from one location to its neighborhood.

Coherence of SAR acquisitions: The calculation of interferograms is very sensitive to the changes that occur between two acquisitions. In particular, sudden changes such as natural disasters, snowfall or new human infrastructures can significantly alter the landscape and thus the properties of the scatterers. Since the signal is not targeting the same scatterers from one acquisition to another, such changes prevent from computing an accurate phase difference in the interferograms. In addition, the layover and shadow areas are inherently incoherent. Since the former result from the backscattered signal from multiple objects, and the latter lack of signal, the displacement in these regions cannot be accurately evaluated.

Although the generation of an interferogram and the detection of ground deformation is theoretically computable, a number of challenges must be overcome. Not only are some regions not reliable at all, but the rest may be affected by external disturbances. Therefore, the interpretation of interferograms requires special attention to avoid inaccurate and erroneous decision making process.

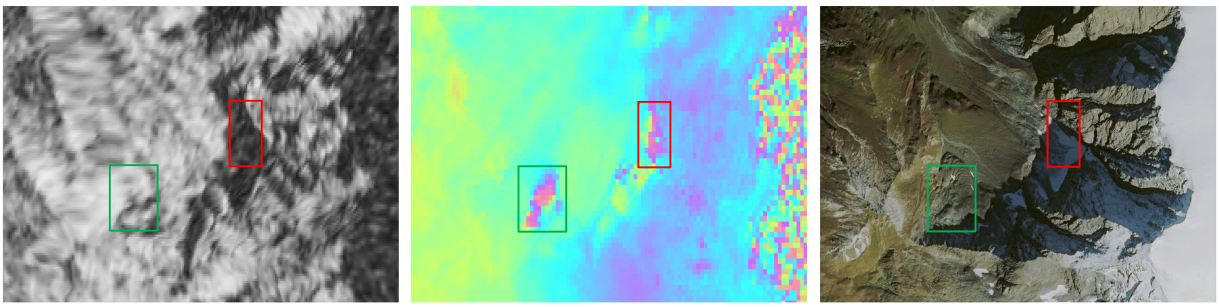
2.4.2 Generating Interferograms

The generation of interferograms from two SAR images can result from diverse algorithms depending on the level of processing and the availability of external data used for refinement. Yet, the mandatory step is the combination of the complex coregistered SAR images using Equation (2.2). For more reliable estimates, the interferogram is computed by the so-called complex multilooking over a neighborhood N:

$$\rho e^{i\Delta\varphi} = \frac{\frac{1}{N}\sum_{n\in N} z_1^n \times \bar{z}_2^n}{\sqrt{\frac{1}{N}\sum_{n\in N}|z_1^n|^2 \times \frac{1}{N}\sum_{n\in N}|z_2^n|^2}} = \frac{\sum_{n\in N} A_1^n A_2^n e^{i(\varphi_1^n - \varphi_2^n)}}{\sqrt{\frac{1}{N}\sum_{n\in N}(A_1^n)^2 \times \frac{1}{N}\sum_{n\in N}(A_2^n)^2}} \tag{2.2}$$

where $z_k = A_k e^{i\varphi_k}$ is the kth image of amplitude A_k and phase φ_k. The notations ρ and $\Delta\varphi$ represent the coherence and phase difference of the interferogram, respectively.

Figure 2.13 displays typical coherence and phase difference maps obtained by such a standard algorithm. In particular, this figure illustrates the importance of the coherence map for the interpretation of the phase difference. Indeed, both the red and green squared regions in Figure 2.13b could be identified as displacement patterns. Nevertheless, due to the low coherence in the red region of Figure 2.13a, its corresponding phase



(a) Interferogram coherence ρ (b) Interferogram phase difference $\Delta\varphi$ (c) Optical image

Figure 2.13: Typical interferogram products (a-b) from Sentinel-1 images on August 19, 2018 and August 25, 2018, and the corresponding SPOT6/7 optical image (c) in 2018 combined by the IGN. The observed rock glacier is located nearby the Col du Vallonnet, France.

difference is not interpretable at all, which means that neither the presence nor the absence of displacement in this region can be stated.

A wide variety of pre- and post-processing can be applied to denoise and account for precise external parameter estimates (*e.g.* atmospheric and topographic delays) for the computation of the interferograms as in [Thollard et al., 2021] and [Lazecký et al., 2020]. For the sake of easy reproducibility, low requirements and fast computation, a standard but effective interferogram pipeline is detailed in Section 6.2.1 and used for slope instability detection in Chapter 6.

2.4.3 Applying to Slope Instabilities

Since SAR images can be acquired regardless of weather conditions and InSAR products can reveal centimeter-scale displacements, this modality has been widely studied for slope instability monitoring, including velocity estimation, collapse precursor detection and susceptibility mapping [Lissak et al., 2020, Mondini et al., 2021, Lacroix et al., 2022].

The InSAR images are widely used to detect slope instabilities over a large region, either before or after an event [Wasowski and Bovenga, 2022]. In particular, in [Bertone et al., 2022], the contribution consists in assessing the kinematics of rock glaciers to complete an inventory covering 11 regions worldwide. Similarly the authors of [Rouyet et al., 2021] leverage multi-temporal InSAR to add kinematics and update previous inventories of rock glaciers and rockslides. The author of [Bodin, 2020] identified 313 rock glaciers in the Vanoise massif in the French Alps using interferograms with a temporal baseline ranging from 6 to 366 days. The study proposed in [Burrows et al., 2019] aims to detect slope instability failures by exploring the coherence map. The main idea is to associate incoherent areas with slope failures using an advanced algorithm. Among other processing steps, the algorithm avoids false alarms due to the presence of vegetation or geometrical distortions. Interferometry was also used to monitor the La Clapière landslide (see Figure 2.2a) in [Fruneau et al., 1996] to estimate its velocity from five images over two weeks.

A more recent branch of interferometry, called PS-InSAR, consists of measuring the displacement only at a few selected points, called persistent scatterers. These scatterers are initially selected as the locations that repeatedly achieve high coherence values over a large stack of interferograms. PS-InSAR allows a more reliable but discrete estimation of the displacement. In particular, it is applied to study the behavior of landslide deformation in the northern Apennines in Italy by installing corner reflectors in [Bayer et al., 2017]. These corners act as artificial point scatterers on the instability surface to ensure the monitoring of a specific constant location. Similarly, PS-InSAR is used in [Intrieri et al., 2019] as an early warning system to find precursors of an open pit mine collapse from Sentinel-1 images. These images are also exploited in [Intrieri et al., 2017] to find precursors of the giant Maoxian mudflow that occurred in the village of Xinmo in China. The study concludes by finding evidence that could have been used to evacuate the population before the collapse. Finally, the authors in [Aslan et al., 2020] and [Bonì et al., 2018] used PS-InSAR to map

landslides in the French and Italian Alps, respectively, and to create inventories of slow-moving areas.

2.5 Conclusion

In this Chapter, an overall context of the current research is developed. In particular, general guidelines on the object of study, slope instabilities, are presented. Special attention is given to the introduction of the different physical and machine learning models that provide decisions depending on the factors that can cause the collapse of slope instabilities. To ensure the implementation of methods that could be regularly updated with frequently acquired data and that could be generalized at a regional scale, remote sensing images are introduced as a key component for slope instability monitoring. Indeed, both optical and SAR images can be used for pre- and post-event inventories. They can also be used to monitor and detect precursory movements by means of optical correlation or SAR interferometry. However, to the best of our knowledge, very few standard multimodal image processing techniques have been applied to slope instability monitoring. One of the rare examples is proposed in [Bodin et al., 2018] where the authors leverage both optical image correlation and InSAR products to map rock glacier displacements in the Chilean Andes. Nevertheless, in the field of deep learning algorithms, multimodal approaches have been widely explored in the last decades, including for slope instability monitoring. Indeed, this Chapter has focused mainly on standard image processing techniques to better understand the advantages and drawbacks of each modality. However, the huge amount of data acquired by satellites makes deep learning approaches a promising approach to solve remote sensing issues. Slope instability monitoring is no exception, and despite the lack of annotated data, a large number of works have explored this topic using deep learning models as reviewed in the next Chapter 3.

Chapter 3

Deep Learning Context

3.1 Introduction

Following the exponential development of deep learning approaches in computer vision, and especially since the creation of Convolutional Neural Networks (CNN), a large number of studies have exploited their learning capabilities for remote sensing image processing. Section 3.2 gives a high-level description of the concept of neural networks and clarifies the terminology used in this thesis. It also describes the major events and the most influential works to better understand the evolution of deep learning algorithms applied to image processing. Then, due to the scope of the thesis, the Chapter focuses on three different research areas and their extension to disaster monitoring applications, in particular slope instabilities. Based on the constraints of both optical and SAR images, Section 3.3 proposes a review of modality translation approaches applied to computer vision and remote sensing images. Following this review, Section 3.4 details numerous approaches for class identification and distinguishes classification from segmentation tasks. Finally, due to the strong requirements for reliability, trustworthiness and understandability of disaster monitoring approaches, Section 3.5 gives insight into the methods developed to explain the "black box" deep learning models. A special focus is put on Class Activation Maps (CAM) to better meet the class identification tasks and to exploit their localization ability.

3.2 Deep Learning for Image Processing

This Section aims to briefly describe the main contributions in deep learning with a special focus on the algorithms implemented for image processing. First, a general context, definitions and the terminology used in this thesis are detailed to avoid confusions caused by the wide variety of vocabulary in the field. Then, the floor is given to the typical architectures from the literature that have inspired a wide range of approaches for image processing. Finally, due to the peculiarity of the works carried out in this thesis, further details are given on the methods that input and output images - also known as image-to-image neural networks.

3.2.1 Overall Context

Deep learning approaches are being increasingly exploited due to the exponential development of the computational capabilities of computers - and especially Graphics Processing Units (GPU). Yet, the history of deep learning goes back to 1962 with the development of the first *perceptron* [Rosenblatt, 1962], *i.e.* an artificial neuron. The aim of this Section is not to trace the whole history of neurons, but only to give the prerequisites for understanding the thesis. For more information on the details of deep learning, the reader is referred to [Nielsen, 2015] where a deep and comprehensible description of the topic is given.

3.2.1.1 Neurons

The concept of a neuron is mathematically simple. The basic perceptron designed in [Rosenblatt, 1962] consists of a linear combination of the inputs as given in Equation (3.1) for an ensemble $\mathcal{X} = \{x_i, i \in N\}$ of N inputs.

$$f(\mathcal{X}) = \sum_{i \in N} w_i x_i + b \tag{3.1}$$

where w_i and b are the weights and the bias of the neuron, known as its *parameters*. The parallel application of several neurons to the same input data forms a *layer* of neurons. Then, by cascading the layers, *i.e.* the output of a given layer l is the input of the following layer $l + 1$, the algorithm becomes a Multi-Layer Perceptron (MLP), also known as a neural network. Such architectures are then composed of an input layer, several *hidden layers*, which are the layers responsible for extracting the most relevant features from the input data, and an output layer.

Although such MLPs can be applied to images by taking each pixel as a separate input, this would severely limit the ability of the algorithms to exploit the spatial and structured aspect of the images. Therefore, the authors of [Lecun et al., 1998] propose the use of *convolutional kernels* instead of perceptrons. Such a kernel $\mathcal{K}$ consists of a fixed number of weights, often organized to form a square matrix. Then, to maintain the spatial consistency of the images before and after the application of the neural function, the authors propose to convolve the input image $\mathcal{I}$ with the kernel. This operation results in an image as the output of the function given in Equation (3.2).

$$f(\mathcal{I}) = \mathcal{I} * \mathcal{K} + b \tag{3.2}$$

where b is the bias of the kernel. In addition, depending on the desire to subsample the features in the same time, one can set a *stride* which controls the overlap of the sliding window used in the convolutional process. Similar to perceptrons, parallelizing multiple convolutional kernels creates convolutional layers (Conv2D) and a series of Conv2D creates a Convolutional Neural Network (CNN).

3.2.1.2 Layers

Depending on the architecture and expected properties of a deep neural network, additional layers can be added. Two of them are of particular interest for this work: *normalization* layers and *activation* layers.

On the one hand, normalization layers aim at facilitating the convergence of the network optimization during the training by learning the most appropriate scaling of the output of the neural layers. The most famous *Batch Normalization* (BN) introduced in [Ioffe and Szegedy, 2015] consists in a normalization based on an estimate of the mean μ and the standard deviation σ of the input of the layer, and on a learned rescaling ensured by two parameters α and β. According to these notations, Equation (3.3) gives the calculation of the normalized features f_{out} with respect to the input features f_{in}.

$$f_{out} = \alpha \frac{f_{in} - \mu}{\sigma} + \beta \tag{3.3}$$

On the other hand, activation layers focus on strengthening the relevance of the activations of neural layers. The term *activation* refers both to the output of a standard neural layer and to the functions used *a posteriori* of these layers. A wide range of activation layers/functions have been implemented to address very specific issues. In this thesis three of them are of particular interest: the Rectified Linear Unit (ReLU) [Nair and Hinton, 2010], the Sigmoid [Goodfellow et al., 2016] and the Softmax [Goodfellow et al., 2016]. The ReLU function is used mainly in the hidden layers to force the network to focus only on the positive activations of the neural layers. Indeed, based on Equation (3.4), all negative activations are set to 0 while the others are left unchanged.

$$ReLU(f_{in}) = \begin{cases} 0 & \text{if } f_{in} < 0 \\ f_{in} & \text{else} \end{cases} \tag{3.4}$$

The other two functions are used more specifically after the very last layer of the network. The Sigmoid function given in Equation (3.5) forces the network to output a value between 0 and 1.

$$Sigmoid(f_{in}) = \frac{1}{1 + e^{-f_{in}}} \tag{3.5}$$

Such a function is widely used for multi-label classification, *i.e.* assigning multiple classes to the same input, since a threshold of 0.5 allows to discriminate whether the class is detected or not. It also serves for most regression tasks, if the expected output is normalized in the range [0,1]. Conversely, the Softmax function is dedicated to mono-labeling (one single label for each input). The latter aims to compare the output activations of the final layer and compute the dominance of one given activation over another. In the case of classification, since each activation is associated with a class, the function helps to evaluate which class the input data are most likely to belong to. The Softmax function is computed as given in Equation (3.6) for the kth output value f_{in}^k among N.

$$Softmax(f_{in}^k) = \frac{exp(f_{in}^k)}{\sum_{j=0}^{N} exp(f_{in}^j)} \tag{3.6}$$

where exp is the exponential function.

Therefore, by combining and alternating these normalization and activation layers with the neural layers, one can build a wide variety of neural networks and ensure their training with backpropagation and stochastic gradient descent.

3.2.1.3 Training

Since neural networks are *learning* algorithms, they require a significant number of training samples to be trained. Depending on the availability of a ground truth associated with each sample, the training is referred to as *supervised* or *unsupervised*. Some methods also exploit alternative ground truths for *weakly supervised* learning or mix supervised and unsupervised datasets for *semi-supervised* learning. Typically, a dataset consisting of annotated images with a class label can be used for the supervised training of a classification network, but only for the weakly supervised training of a segmentation network (more details on the distinction between classification and segmentation are given in Section 3.4). Conversely, a dataset containing only input images can be used for an unsupervised training, or it can be coupled with an annotated dataset to perform a semi-supervised training.

The training of a neural network is a four-step iterative process aimed at updating the learnable parameters of the network, *i.e.* the weights and biases of the neural layers, and the normalization parameters α and β of the BN layers.

Forward pass: The forward pass consists of passing an input to the network, *i.e.* applying each successive layer to the input to obtain the final network decision. Yet, since the learnable parameters are initialized randomly, the output of the network is necessarily random at first.

Loss calculation: Based on the output obtained in the previous step, the objective is to measure or estimate the error made by the network with respect to its expectation. To this end, a *loss function* is used to compute the error. The network is then trained to minimize this error and to obtain the most accurate performance with respect to this criterion. In the case of supervised training, the loss measures a distance between the output of the network and the ground truth associated with the input data.

Backward pass: Since the objective of the network is to minimize the loss function, this basically means that the partial derivatives of this function with respect to each parameter of the network should be 0. Therefore, the third step is to compute the gradient of the loss function. The term *backward pass* refers to the way in which the gradient is computed. Indeed, from a computational point of view, it is much more efficient to compute the gradient of the late parameters first, and backpropagating this gradient through the network until the input [Rumelhart et al., 1986]. Since the description in this Section is meant to be high-level, no further details are given, but a comprehensive description of the process can be found in [Nielsen, 2015]. Ideally, if the loss is a convex function, achieving a zero gradient means that the performance of the

network reaches the global minimum error. But in practice, due to the complexity of neural networks, it is more likely that they reach a local minimum, *i.e.* the algorithm is perfectible.

Optimization: The final step that allows the network to learn from the data is the update of its parameters. The standard algorithm for such an operation is based on the Stochastic Gradient Descent (SGD) [Goodfellow et al., 2016] which consists in modifying the value of each parameter according to the opposite direction of the calculated gradient, as in Equation (3.7).

$$w_{i+1} = w_i - \lambda \frac{\partial L}{\partial w_i} \tag{3.7}$$

where w_i is the value of the parameter at the ith iteration, L is the loss computed during the forward pass and λ is the learning rate to moderate the value of the gradient and avoid divergence. Starting from such an optimizer, a variety of further refined optimizers have been created. In particular, this thesis focuses on the widely used Adam optimizer [Kingma and Ba, 2014] which proposes to consider previous gradients to limit the noise during the learning process.

Due to the high demand of data samples to be used for training the network, not all of them can be sent to the network at once. Therefore, the training is organized in *epochs* which consists in iterating the four previous steps on a random set of samples from the training dataset called *batch*. The stochasticity of the SGD and Adam, and therefore the noise, is due to such batch splitting. From one batch to another, the gradient can guide the training in different ways which requires some regularization as proposed in the Adam optimizer [Kingma and Ba, 2014]. To better distinguish the parameters of the networks from the parameters that the user can select, the latter are called *hyperparameters*. These hyperparameters - including the learning rate, the number of epochs, the batch size and even the architecture of the networks - are specifically studied to find the best combination that achieves the lowest of loss minima without suffering from *overfitting*.

The overfitting is a training artifact that consists in specializing the network to the training dataset without being able to generalize to other datasets. The most critical example of overfitting on a training dataset $\mathcal{T}$ of a network f is given in Equation (3.8).

$$f(x) = \begin{cases} y & \text{if } x \in \mathcal{T} \\ rand & \text{else} \end{cases} \tag{3.8}$$

where x is an input, y is its actual ground truth and $rand$ is a random value. To detect such an artifact during the training, the available dataset is often split into two or three subdatasets:

- A training dataset consisting of most of the samples used to train the parameters of the network.
- An evaluation/validation dataset consisting of fewer samples but independent of the training dataset. Depending on the objective of the proposed methodology, the dataset can be used to prove the generalizability of the network or to fine-tune the hyperparameters.
- A test dataset that is independent of the other two and is necessary to prove the generalizability of the overall method if the evaluation/validation dataset is used for hyperparameter tuning.

Several approaches can be implemented to delay the overfitting of the network: regularization, dropout layers [Hinton et al., 2012], additional loss constraints, or increasing the size of the training dataset with *data augmentation* strategies. The latter consists in artificially creating new samples by transforming the actual dataset samples. Such transformations include the addition of noise, geometric transformations (*e.g.* translations, rotations, flips), compression, but also more sophisticated strategies such as MixUp [Zhang et al., 2018a] or CutMix [Yun et al., 2019].

3.2.2 From U-Net to Transformers

The availability of very large datasets such as ImageNet [Deng et al., 2009], PASCAL VOC [Everingham et al., 2010], MNIST [Lecun et al., 1998] and CityScape [Cordts et al., 2016] encourages for the implementation of large neural networks and novel layers while limiting the susceptibility to overfitting.

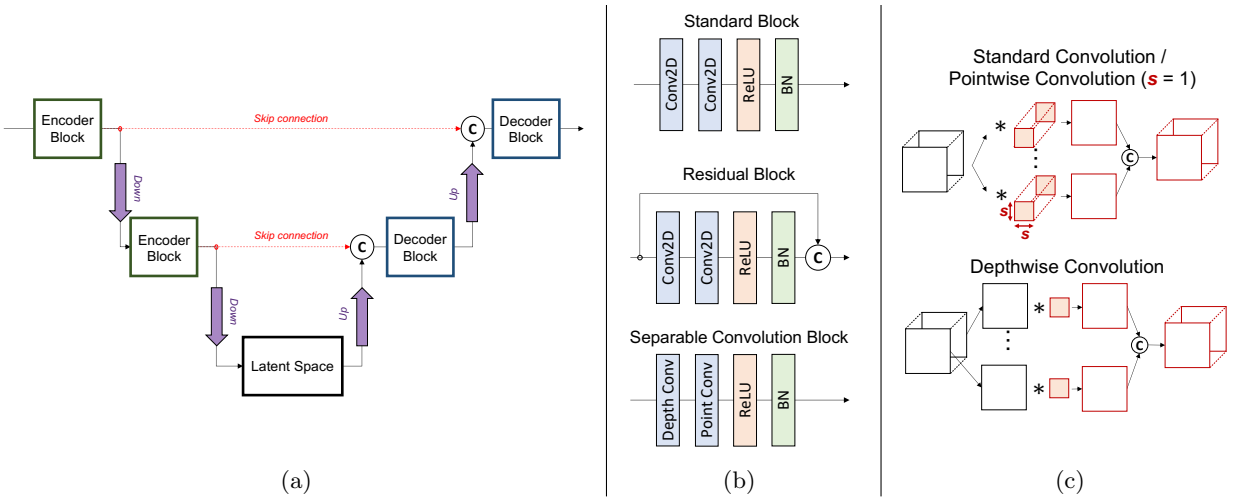


Figure 3.1: Typical deep learning breakthrough from the literature. (a) refers to an example of the U-Net architecture [Ronneberger et al., 2015]. (b) gives the example of a standard, residual [He et al., 2016a] and separable convolution [Chollet, 2017] block. (c) illustrates the difference between standard, pointwise and depthwise convolutions. The circled C in the different schemes refers to the concatenation of the feature maps along the channel dimension.

3.2.2.1 U-Net and its derivatives

One of the keystones in the field of deep learning applied to images is the implementation of the U-Net architecture in [Ronneberger et al., 2015]. This architecture given in Figure 3.1a is designed as an *encoder-decoder* architecture. The encoder part takes an input, typically an image, and extracts the most relevant features with respect to that input based on a series of convolutional layers, ReLU activations, Batch Normalization layers and *MaxPooling* subsampling. The latter layer is basically used to reduce the dimension of the features extracted by the previous layers. It subsamples each neighborhood of its input with its maximum value. In this way, the U-Net encoder is able to increase the number of convolutional kernels without exploding the memory capacity of the computer. The output of the encoder, known as *latent space*, is the code containing the most relevant features from the input image (or data) with respect to the target task. This code is further processed to be projected into the latent space of the desired output images. The resulting code is then fed to the U-Net decoder which consists of a series of Conv2D, ReLU, BN and Upsampling or Transposed Convolutions used to progressively increase the size of the reconstructed features. The peculiarity of U-Net architectures compared to standard encoder-decoder models is that *skip connections* are introduced between each stage of the encoder and the decoder. In this way, the authors propose to transfer the high-resolution features lost during the pooling steps directly to the decoder layers to provide more detailed outputs.

This architecture has been widely extended and derived into many variants such as U-Net++ [Zhou et al., 2018] which integrates more skip connections between the different blocks. The standard convolutional blocks themselves have been replaced by *Residual Blocks* [He et al., 2016a] or *Separable Convolutions* [Szegedy et al., 2016, Chollet, 2017] illustrated in Figure 3.1b. On the one hand, the residual blocks consist of the integration of a connection between the input of the series of convolutional layers and their output. Thus, the vanishing gradient issue, which is known to negatively affect the training of very deep networks, can be limited and their optimization can converge faster. On the other hand, separable convolutions as described in [Chollet, 2017] are based on the separation of spatial and inter-channel feature extraction. As visible at the bottom of Figure 3.1c, a depthwise convolution is used to apply a different convolutional kernel to each channel independently and extract spatial features. Then, the pointwise convolution on top of Figure 3.1c, *i.e.* the linear combination of the channels, is applied to extract the inter-channel features. Although U-Net is a true breakthrough in deep learning for image processing, its variants have demonstrated strong capabilities to enhance its performance in a variety of image-based applications.

3.2.2.2 Multi-source inputs

The U-Net architecture can be limited in terms of interaction between inputs of different modalities, sources and types. Therefore, three common architectural tricks are often implemented to ensure a more relevant data interaction: *early fusion*, *intermediate fusion* and *late fusion* as experimented in [Garnot et al., 2021].

The early fusion consists of fusing the two data into a single input at the beginning of the network. In the case of different images, it is often based on a concatenation along the channel dimension, while in the case of data with different structural types, an embedding *a priori* of the network is applied to merge both inputs.

The intermediate fusion first applies a series of layers to each input separately before merging them at some point in the network, often in the latent space. The rest of the network is then applied to the merged features before making the final decision.

The late fusion is based on independent processing of each input throughout the network. The features are then fused in the final layers before producing the decision. Some approaches known as *decision fusion* even propose to merge the output of each independent network and avoid any interaction between the inputs.

The main characteristic of these strategies is that they assume that both inputs are equally relevant to the decision task. However, some data could be provided to the user as an additional input to better guide the network, avoid errors or provide low-level information. Therefore, giving the same importance to each input during the training may limit the learning abilities of the network. To enforce the dominance of one input over another, the authors of [de Vries et al., 2017, Dumoulin et al., 2017, Mescheder et al., 2019] propose the use of Conditioned Batch Normalization (CBN). The latter consists of embedding a conditioning input into a latent code and using this latent code to modify the α and β parameters of the standard BN layers in Equation (3.3). This modification helps to guide the behavior of the main network responsible for processing the most informative input.

Further works aim at a common representation of different input data. In this way, two data from different sources can be represented with the same latent code and processed in the same way. The authors of [Ganin and Lempitsky, 2015] propose to use an adversarial training (described in more detail in Section 3.3.1) to force a single feature extractor to find common features between two input data acquired from different sources. Another approach proposed in [Jain et al., 2021] is to learn the same representation based on the BYOL [Grill et al., 2020] or the SimSiam [Chen and He, 2021] method. The main advantage of this approach is its self-supervised learning capability, *i.e.* it is trained without any target task except finding a common latent representation.

3.2.2.3 Generative networks

Besides encoding the most characteristic features of an input data, neural networks are also widely explored for generation purposes. The idea behind these algorithms is to learn and reproduce the specific features from a given distribution. In this way, starting from a random input, the neural network is able to generate a completely artificial sample from the distribution of interest. These approaches can be divided into three different strategies: Generative Adversarial Networks (GAN), Variational Auto-Encoders (VAE) and diffusion models.

The implementation of GANs in [Goodfellow et al., 2016] is particularly exploited because they have strong hallucination capabilities, *i.e.* capabilities to generate credible features without *a priori* knowledge. Indeed, the main generator network aims to produce an output image that contains the same specific patterns as the images from a given output distribution. Its goal is to fool an auxiliary classifier network, called a discriminator, which is trained to classify whether the input image is from the real or the generated distribution. Since these architectures are widely used and derived for modality translation, a more detailed description is provided in the dedicated Section 3.3.1.

For a long time, the main competitors of GANs were the VAEs introduced in [Kingma and Welling, 2014]. This strategy is based on a standard auto-encoder architecture, *e.g.* a U-Net without skip connections. The peculiarity of the VAE is that the encoder does not aim to extract a "latent code" but the latent statistical properties of the input image. This latent space consists of a mean vector μ and a standard deviation vector σ of a given size N. Based on these statistical properties, a new vector is randomly generated where each

component c is sampled from a Gaussian distribution with its respective mean μ_c and standard deviation σ_c from the latent vectors. The decoder is then trained to reconstruct the input image based on the sampled vector. At inference time, the VAE consists of extracting the latent statistical vectors from a given image of interest and generating multiple samples with the same statistical properties.

Yet, with the increase in computational power, another generative approach is gaining visibility. Diffusion models proposed in [Sohl-Dickstein et al., 2015] consist of iterative processes that aim to progressively denoise an input until a clean output is generated. For this purpose, an input image is first iteratively noised T times. Then, a neural network is trained to denoise each step $t \in [1, T]$ so that the output of the network is the image at step $t-1$. Thus, an image with random pixel values can be generated at inference time, and by iteratively applying the network to this noisy image, a clean image can be generated. Following the rapid development of diffusion models, the authors of [Dhariwal and Nichol, 2021] show that they are highly competitive with GANs when asked to generate images similar to ImageNet [Deng et al., 2009].

3.2.2.4 Multi-temporal inputs

Focusing on the properties of input data, adequatly addressing their temporality is also a major concern. Both computer vision videos and stacks of remote sensing images of the same location are similar in terms of temporal consistency. However, such temporality is not or poorly taken into account by standard CNNs and MLPs.

Therefore, Recurrent Neural Networks (RNN) are proposed to better exploit the temporal consistency of the input data. The authors of [Hochreiter and Schmidhuber, 1997] propose the Long-Short Term Memory (LSTM) units which is a reference in the field. These units are originally implemented for the Natural Language Processing (NLP) task by taking each embedded word successively as input. The LSTM uses each input to update a long-term memory unit, which extracts the meaning of the whole sentence, and a short-term memory unit, which finds out the local context of each word. This strategy is further refined into Gated Recurrent Units (GRU) in [Cho et al., 2014] where the long-term memory unit is removed and the update of the short-term memory unit is guided by an *a priori* estimation of the relevance of each word. The combination of these works with convolutional layers results in ConvLSTM networks [Shi et al., 2015] and ConvGRU networks [Ienco et al., 2019]. Although both show competitive performance when applied to temporal image stacks, they lack stability in the training. Therefore, Temporal Convolutional Networks (TCN) are proposed in [Lea et al., 2016] to avoid the use of RNN but still exploit the temporality of the images by using 1D convolutions. A similar idea is applied in [Garnot et al., 2021] after embedding radar and optical images in a similar latent space.

3.2.2.5 Transformers

To mitigate the unstable performance of RNNs applied to NLP tasks, another solution is to leverage attention mechanisms. Attention mechanisms basically compute an attention map that evaluates the relationship between two pieces of the same input data. This attention map can then be processed to make a decision or used to force the network to focus on the most related components of the data. The parallelization of such attention mechanisms in multi-head attention module layers and the cascading of such multi-head layers leads to a new type of neural networks: the *Transformers* [Vaswani et al., 2017]. Although this architecture is originally dedicated to NLP algorithms, it has been rapidly extended to image processing with Visual Transformers (ViT) in [Dosovitskiy et al., 2020]. Such image-specific transformers lead to more complex and powerful architectures such as SwinTransformers [Liu et al., 2021b], which are designed to perform hierarchical feature extraction based on a transformer network.

Despite the fact that these approaches are already widely used in remote sensing [Kandala et al., 2022, Orlandi et al., 2022, Wang et al., 2022d], they are not explored in this thesis for two main reasons. First, transformer architectures are particularly heavy networks that require time, memory, powerful machines and a huge amount of data to be accurately trained. The latter aspect is extremely limiting in the context of this thesis, *i.e.* the supervised training of a network on real annotated images. Second, the works which are conducted in this thesis aims to be easily understandable by non-experts in the field. Despite the

increasing interest in transformers, CNNs have been widely studied in terms of explainability, as described in Section 3.5, which makes them a better choice for the dissemination of these works.

3.2.3 Image-to-Image Applications

The implementation of the aforementioned network architectures is widely leveraged for many image processing applications. Although image-to-scalar applications such as classification or physical parameter estimation also rely on such architectures, this thesis focuses on image-to-image strategies. Such approaches preserve the spatial aspect of the input data and provide a pixel-wise estimation of the result, which is particularly informative in the case of slope instabilities.

The first class of image-to-image algorithms addresses the correction of the artifacts in an input image. Noise, blur, pixelation, missing data, or unwanted objects are so many artifacts that can be mitigated by using deep learning models. Although the denoising issue is mostly tackled with supervised encoder-decoder networks, it is more challenging when no ground truth is available. To address this issue, the authors of [Lehtinen et al., 2018] propose to leverage multiple acquisitions of the same image under different configurations. The encoder-decoder architecture is then trained to make the reconstruction of one noisy input image similar to another. Since the network aims to reconstruct the general aspect of the image, it succeeds in denoising the image despite the noisy ground truth. Even less requirements are proposed by the authors of [Krull et al., 2019] and [Batson and Royer, 2019] as their strategy is to make the reconstruction matching the input image or its neighborhood. Inspired by such performance and the non-necessity of actual ground truth, this strategy is also extended to the despeckling of SAR images in [Dalsasso et al., 2021].

The opposite issue, *i.e.* the lack of contrast due to pixelation or blurriness, is often addressed based on decoder-only networks to increase image resolution. These so-called super-resolution methods aim at guessing the most susceptible interpolation to be applied to a pixel given its neighborhood. The authors of [Shi et al., 2016] propose a novel upsampling layer called sub-pixel convolutional layer to perform the super-resolution. Such a layer consists of applying a certain number of convolutional kernels to the low-resolution images and organizing each pixel of the feature map to reconstruct the final upsampled image. More details about this strategy are given in Section 4.4.2. This approach is improved by the authors of [Aitken et al., 2017] through a specific initialization of the convolutional weights that reduces the checkerboard artifact observed in the reconstructed images. On the other hand, the authors of [Li et al., 2019] propose the feedback block, which consists in iteratively applying the same block to an input. In this way, the block is forced to leverage its own output to reconstruct more details in the image and thus produce a more accurate super-resolution image. Due to the low resolution of the vast majority of remote sensing images, such super-resolution methods are also widely explored in the literature as illustrated by the review in [Wang et al., 2022c].

Another critical issue in image processing is the missing data due to no data values or to the presence of an object that hides the backside information, *e.g.* the clouds on optical images. The various approaches to reconstruct the missing information are often referred to as inpainting strategies and are typically addressed with iterative optimization methods [Cheng et al., 2017]. Yet, deep learning approaches have shown excellent performance in reconstructing the most susceptible underlying patterns. For instance, the authors of [Imran et al., 2022] propose to combine deep convolutional layers with a robust Principal Component Analysis (PCA) supported by the cloud mask to perform cloud removal. Two other approaches in [Bermudez et al., 2018] and [Darbaghshahi et al., 2022] are based on modality translation to exploit the SAR cloud penetration capabilities and produce a cloud-free optical image.

Such modality translation methods, together with style transfer and even image segmentation do not aim at image correction but at image transformation. The image segmentation, and especially semantic segmentation, focuses on transforming the image into different masks associated with classes of interest. Contrary to this discrete transformation, style transfer and modality translation are even more challenging. Indeed, these two transformations are often ill-posed, since a given input can refer to several different outputs. Modality translation and, to some extent, style transfer are described in more detail in the following Section 3.3 to better understand the pros and cons of such approaches. Similarly, the methods addressing image segmentation and more generally the identification of classes in images are the core of Section 3.4.

3.2.4 Performance Evaluation

The evaluation of the performance of neural networks is usually based on two criteria: qualitative inspections and quantitative metrics. While the former result from a subjective evaluation, the latter are based on actual mathematical formulas. In the following paragraphs, the metrics used in this thesis are described in detail, depending on their use for classification/segmentation or regression evaluation.

3.2.4.1 Classification and Segmentation Metrics

The calculation of most classification and segmentation metrics is based on the confusion matrix as given in Equation (3.9) when considering a binary task, *i.e.* the output label is either 0 or 1.

$$M = \begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix} \tag{3.9}$$

where TN and TP are the True Negatives and True Positives, and FN and FP are the False Negatives and False Positives. The number of TN and TP correspond to the correct detections of the network while the number of FN refers to the samples falsely labeled as 0 and the number of FP refers to those wrongly labeled as 1. Depending on the situation, one may trade off decreasing the FN against increasing the FP, as in disease detection and in our particular case of slope instability detection, or one may prefer decreasing the FP over the FN, as in judicial decision making. A number of metrics can be derived from these TP, TN, FP and FN.

Accuracy (Acc.): The accuracy, also called overall accuracy, is the ratio of true decisions to the total number of samples as in Equation (3.10). It is used to assess the global performance of the network but its calculation can be a source of misinterpretation, especially if the dataset suffers from a strong class imbalance.

$$Acc. = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.10}$$

Precision (Prec.): The precision evaluates the ability of the network to make an accurate prediction, *i.e.* if the class 1 is detected, it really belongs to class 1. Yet, it does not take into account the number of false negatives in the process as visible in Equation (3.11). This means that the precision does not inform about the ability of the network to find all actual samples of class 1.

$$Prec. = \frac{TP}{TP + FP} \tag{3.11}$$

Recall (Rec.): The recall is the complement of the precision since it evaluates the ability of the network to detect the actual class of interest, *i.e.* if a sample belongs to class 1, it is accurately detected as class 1. Unlike precision, the recall does not take into account the number of false positives in its calculation, which is given in Equation (3.12). This means that the recall does not guarantee that all predictions are correct, which shows the complementarity of the precision and the recall metrics.

$$Rec. = \frac{TP}{TP + FN} \tag{3.12}$$

***F1-Score (F1-S)* / Dice**: The F1-Score or Dice is calculated as the harmonic mean of the precision and the recall. It helps to find out the actual overall performance of the network. Yet, it does not provide insight into the limitations of the classifier since a low F1-Score value can be caused by a low recall, a low precision or both. Its calculation is given in Equation (3.13).

$$F1\text{-}S = \frac{2 \times Rec. \times Prec.}{Rec. + Prec.} \tag{3.13}$$

It is worth noting that in the case of segmentation, the F1-Score does not reward an accurate zero map prediction, *i.e.* when the class of interest is not in the image. Indeed, since there are no TP to be found,

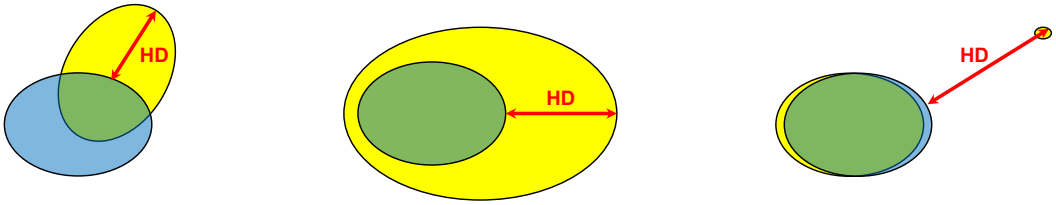


Figure 3.2: Hausdorff Distance (HD) in three different cases where the blue ensemble is the ground truth, the yellow ensemble is the prediction and the green region is the overlapping (true positive) region. From left to right, ensembles are overlapping, one ensemble is included in the other, and the presence of an outlier as part of the predicted ensemble.

both the precision and the recall are necessarily equal to 0, which lowers the F1-Score even if the network achieves an accurate zero segmentation map. Therefore, the F1-Score is often extended to the Dice Score, *e.g.* in the *miseval* library[1], to take such predictions into account in the calculation by setting the F1-Score to 1 in these particular cases.

Area Under the Curve (AUC): This metric measures the area under the Receiver Operating Curve (ROC). The ROC curve is a function that plots the evolution of the False Positive Rate (the precision of class 0) against that of the True Positive Rate (the precision of class 1) depending on the threshold used for binary decision. Due to the nature of neural networks, the final output score is not a binary label but a value between 0 and 1. Therefore, the performance of the network can vary drastically depending on the threshold chosen to compute the TP, TN, FP and FN mentioned above, whereas the AUC evaluates the discriminability of the classifier.

mean Intersection over Union (mIoU): This metric is considered as an ensemble metric and is most commonly used in segmentation. It measures the ratio between the overlap of the ground truth and the predictions compared to their union. Basically, by considering A as the ground truth annotation and P as the binary map predicted by the network, the mIoU is given by Equation (3.14).

$$mIoU = \frac{A \cap P}{A \cup P} \tag{3.14}$$

Hausdorff Distance (HD): This final segmentation metric is used to compute the maximum projected distance of the boundaries of the predicted segmentation onto the actual annotation. Figure 3.2 gives a visual explanation of the measured distance. The closer the ensembles are, the lower the metric. It is particularly relevant for evaluating the presence of outliers in the results.

3.2.4.2 Regression Metrics

Unlike classification and segmentation tasks, the regression task requires the evaluation of a real and continuous output. Therefore, measuring the performance of a neural network cannot be done with discrete metrics. In this thesis, five metrics are used to evaluate regression performance. It is worth mentioning that the thesis focuses on images. Therefore, the interpretation of the metrics given below considers the particular context of a regression task based on image outputs.

Mean Absolute Error (MAE): The mean absolute error computes the L1 distance between the pixel values of the predicted image $\hat{\mathcal{I}}$ and the expected true image $\mathcal{I}$. The MAE calculation straightforwardly averages the different errors as given in Equation 3.15.

$$MAE = \frac{1}{H \times W} \sum_{(i,j) \in H \times W} |\hat{\mathcal{I}}_{i,j} - \mathcal{I}_{i,j}| \tag{3.15}$$

[1] https://github.com/frankkramer-lab/miseval, last visited on July 13, 2024

where H and W refer to the height and width of the image. This metric is mostly used to measure the overall differences (*e.g.* in terms of brightness and colors) between the two images while reducing the effect of outlier pixels.

Mean Square Error (MSE): Similar to the MAE, the MSE replaces the use of the L1 distance with the L2 distance to compare the two images as in Equation (3.16). Although it is more sensitive to the effect of outliers, it remains a strong indicator of the overall quality of the generated images.

$$MSE = \frac{1}{H \times W} \sum_{(i,j) \in H \times W} (\hat{\mathcal{I}}_{i,j} - \mathcal{I}_{i,j})^2 \tag{3.16}$$

To get a better idea of the true distance between the true and predicted images, the Root Mean Square Error (RMSE) is often calculated as $RMSE = \sqrt{MSE}$.

Structural Similarity Index Measure (SSIM): The SSIM measures the local similarity between two images. Given a window of size N, the SSIM computes the similarity of the distributions of the neighborhood of each predicted pixel with respect to the true image. Such a metric gives a better insight into the structurality of the different reconstructed patterns than the MSE and MAE, since it is computed locally. Equation (3.17) gives the computation of the SSIM score by considering σ_x^2 as the variance, $\sigma_{x,y}$ as the covariance, μ_x as the mean, and ϵ_i as a small coefficient avoiding division by 0. All statistics are computed over the neighborhood of the pixel given by the window size N. To avoid heavy notation, this information is omitted in Equation 3.17, and $\hat{\mathcal{I}}$ is denoted as I.

$$SSIM = \frac{(2\mu_I \mu_{\mathcal{I}} + \epsilon_1)(2\sigma_{I,\mathcal{I}} + \epsilon_2)}{(\mu_I^2 + \mu_{\mathcal{I}}^2 + \epsilon_1)(\sigma_I^2 + \sigma_{\mathcal{I}}^2 + \epsilon_2)} \tag{3.17}$$

Peak Signal to Noise Ratio (PSNR): The PSNR evaluates the quality of the reconstructed image based on its noise level compared to the original image. Depending on the encoding of the image, the PSNR is calculated based on the maximum dynamic d of the image (typically 1 if encoded in $[0, 1]$ or 255 if encoded in $[0, 255]$). Equation (3.18) is then used to compute the PSNR.

$$PSNR = 10log(\frac{d^2}{MSE}) \tag{3.18}$$

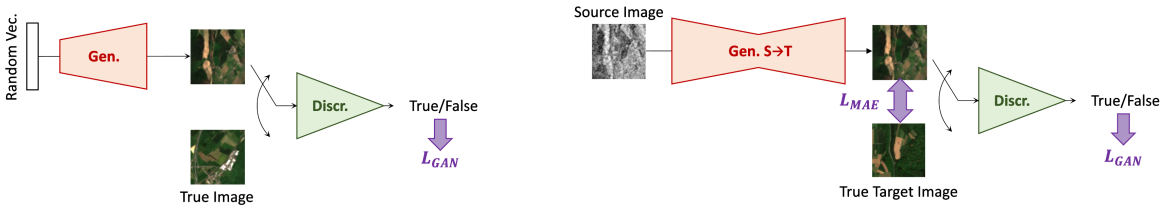
It is worth mentioning that in the current study, the behavior of certain methods can be different when calculating the PSNR and MSE scores. Indeed, the MSE scores are always calculated on the network output values. Conversely, the PSNR scores are computed after thresholding the output values lower than 0 (set to 0) and greater than 1 (set to 1) and encoding them on 8 bits. Therefore, the MSE score is more sensitive to outlier values outside the expected [0,1] range than the PSNR score.

Fréchet Inception Distance (FID): This last metric has a different goal than the others. Instead of measuring the relevance of the patterns, it aims to evaluate their credibility. Thus, it is not computed pixel by pixel or image by image, but directly over the entire dataset. The FID evaluates the difference in the distribution of the features extracted by the Inception network [Szegedy et al., 2016]. Given a layer of interest, all predicted (resp. true) images are passed into Inception and the features X (resp. Y) of that particular layer are examined. It is the comparison of the distributions of these features extracted from the predicted images with the true images that computes the FID score as given in Equation (3.19).

$$FID = ||\mu_X - \mu_Y||^2 + tr(\Sigma_X + \Sigma_Y - 2\sqrt{\Sigma_X \Sigma_Y}) \tag{3.19}$$

where μ_i and Σ_i refer to the multivariate mean and covariance matrices, and $tr()$ refers to the trace operation.

Although the MSE, MAE, SSIM and PSNR are widely used in the literature, the FID score is of great interest when dealing with generative algorithms. However, since this metric only measures the credibility of the final image, it does not take into account the relevance of the reconstructed patterns. Such relevance is a matter of great concern in remote sensing and especially in modality translation approaches since it helps to find out whether the network has successfully transferred the input features into output accurate features.



(a) GAN [Goodfellow et al., 2014]

(b) Pix2pix [Isola et al., 2017]

Figure 3.3: Standard adversarial training of generative networks.

3.3 Modality Translation

Since slope instabilities are mainly triggered by intense rainfall, such conditions imply the presence of dense clouds over the regions of interest. Nevertheless, since the optical images cannot penetrate these clouds, SAR-to-optical translation is the preferred strategy to provide fast decisions for sudden landslide detection but also for floods or forest fires [Luppino et al., 2021, Li et al., 2021b, Zhao et al., 2023]. In addition to these emergency applications, modality translation is explored to improve the quality of several related tasks such as the land-cover classification of SAR images [Ley et al., 2018], optical image registration [Merkle et al., 2017, Merkle et al., 2018], cloud removal [Bermudez et al., 2018, Darbaghshahi et al., 2022] or super-resolution [Feng et al., 2022].

In this study, modality translation and style transfer are considered as two different but related tasks. The former consists in extracting the most characteristic features of an image and mapping them to the features of another modality - *i.e.* transforming an input image into another image with different physical properties. Conversely, style transfer aims to generate an image in the same modality but acquired under a different configuration than the input. As a matter of fact, the SAR-to-optical translation is considered as a modality translation task (from structural to radiometrical properties) but the summer-to-winter translation is considered as a style transfer task (the configuration is the seasonality) as is the SAR VH-to-VV translation (the configuration is the polarization of the received signal).

Modality translator networks are of particular interest in remote sensing to exploit the vast amount of data provided by satellites. Since this task requires only the availability of SAR and optical images to be translated, deep translators can be trained on any multimodal dataset, whether annotated (SpaceNet 6 [Shermeyer et al., 2020], BigEarthNet-MM [Sumbul et al., 2021], DREAM [Koeniguer et al., 2021]) or not (SEN1-2 [Schmitt et al., 2018], SEN12-MS [Schmitt et al., 2019]).

It is worth noting here that the modality translation approaches implemented in Chapter 4 are further used in Chapter 5 for slope instability detection. This downstream change detection requires the reconstruction of actually reliable and relevant patterns. Thus, although the reconstruction of the image are less visually credible, the search for relevant reconstructed pattern is the core of this thesis. Methods for credibility enhancement are only considered here as a bonus for better visual appreciations.

This section first reviews the different strategies implemented in the literature to address modality translation from its applications in computer vision to its extensions to remote sensing images. A special focus is then given to supervised and unsupervised adversarial networks proposed for remote sensing modality translation. Finally, alternative approaches that aim to avoid the drawbacks of an adversarial training are described.

3.3.1 From Computer Vision to Remote Sensing

Image-to-image translation is studied in computer vision not only for generation purposes but also to tackle various challenging tasks, including dehazing [Zhang and Li, 2021], segmentation [Piva and Dubbelman, 2023], super-resolution [Chen et al., 2022a] and image coloring [Iizuka et al., 2016]. Although these applications are more related to the definition of style transfer given above, they use similar strategies to modality

translation. In particular, the core concept of style transfer and modality translation networks is based on Generative Adversarial Networks (GAN) from [Goodfellow et al., 2014].

The original GAN architecture, as depicted in Figure 3.3a, consists of two competing subnetworks: a generator and a discriminator. The generator aims to create an image following a target distribution from a random input vector. To force the generator to produce realistic images, a discriminator network is trained to detect whether an input image comes from the true distribution or from the generated one. Thus, from one epoch to the next, the generator and the discriminator compete to maximize and minimize, respectively, the classification loss applied to the output of the discriminator - typically the cross-entropy loss L_{GAN} given by the Equation (3.20).

$$L_{GAN}(X_1, X_2) = -\frac{1}{\#X_1} \sum_{x_1 \in X_1} log(D(x_1)) - \frac{1}{\#X_2} \sum_{x_2 \in X_2} log(1 - D(x_2)) \tag{3.20}$$

where D is the discriminator ending by a Softmax layer to output the probability of belonging to X_1, and X_1 and X_2 are two ensembles of images from the same modality but with two different distributions.

Based on this loss and by considering the following notations: X as the ensemble of true images and $\hat{Y} = \{G(z),\ \forall z \in Z\}$ as the ensemble of images created by the generator G from the ensemble of random vectors Z sampled during the training epoch, the competition created by the adversarial framework implies that the generator aims to minimize $L_{GAN}(\hat{Y}, X)$ while the discriminator aims to minimize $L_{GAN}(X, \hat{Y})$. These strategy forces the generator to *hallucinate* features to mimic the distribution of the real images, *i.e.* to create patterns which are the most probable to belong to the distribution without any cue about its relevance in the context. This hallucination property is highly sought for generative approaches because it allows to create credible images from a qualitative point of view. The main issue with GANs is that the random vector does not allow to choose the properties of the generated images. Therefore, the authors of [Mirza and Osindero, 2014] propose the conditional GAN (cGAN) which inputs an additional controllable vector to the generator. Since the vector is associated with given characteristics of interest, it forces their creation in the final image.

Inspired by the performance of GANs and cGANs, the authors of Pix2pix [Isola et al., 2017] propose to remove the random vector and to design the generator as an encoder-decoder architecture - namely, U-Net from [Ronneberger et al., 2015]. To this end, the encoder of the generator aims to extract the most relevant features from an input to condition the generation of its corresponding image in another modality by the decoder as visible in Figure 3.3b. This architecture is implemented for both style transfer and modality translation under a supervised configuration - *i.e.* with paired images. This setting allows the calculation of the Mean Absolute Error loss (L_{MAE}) to match the translation of the input image with its actual ground truth image. The authors have found that the MAE results in less blurry reconstruction than the MSE and strengthens the hallucination abilities of the network. Thus, by considering S and T as the paired ensembles of source and target images, the generator of Pix2pix aims to minimize the loss L_{P2P} given by the Equation (3.21), where λ is a hyperparameter weight used to balance the contributions of the different terms.

$$L_{P2P}(S, T) = L_{MAE}(G(S), T) + \lambda L_{GAN}(G(S), T) \tag{3.21}$$

Nevertheless, the pairing is not always available for use. To address this issue, the authors of [Zhu et al., 2017] propose CycleGAN as depicted in Figure 3.4. The network is composed of two generators so that the first, $G_{S \to T}$, translates an image from a source modality S to a target modality T and the second, $G_{T \to S}$, performs the reverse translation. While the loss applied to the classification of the source and target discriminators remains unchanged as in the standard GANs networks, the error computed on the output of the generators cannot be evaluated with standard supervised losses. Therefore, the authors propose a cycle consistency loss which consists in translating an input image from the source to the target modality and then back to the source modality. The source reconstruction is then compared to the original source image with the MAE loss. The same process is applied to the images of the target modality. Mathematically, the cycle consistency loss is calculated by the Equation (3.22).

$$L_{cycle}(S, T) = L_{MAE}(G_{T \to S}(G_{S \to T}(S)), S) \tag{3.22}$$

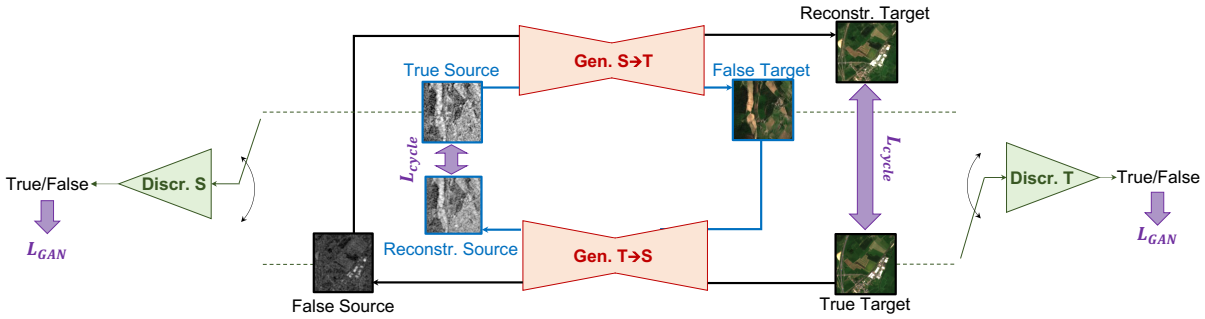


Figure 3.4: Adversarial training of CycleGAN [Zhu et al., 2017]

In addition, to ensure that the generators are acting as an identity function when a target image is given as input, an identity loss term is also exploited through Equation (3.23).

$$L_{idt}(S,T) = L_{MAE}(G_{T\to S}(S), S) \tag{3.23}$$

Combining these losses with the GAN losses from the source and target discriminators D_S and D_T results in the final CycleGAN loss, L_{CG} in Equation (3.24), which is applied to the generators.

$$\begin{aligned} L_{CG}(S,T) = \quad & L_{cycle}(S,T) + L_{cycle}(T,S) \\ & +\lambda_{idt}\ (L_{idt}(S,T) + L_{idt}(T,S)) \\ & +\lambda\ (L_{GAN}^{D_T}(G_{S\to T}(S),T) + L_{GAN}^{D_S}(G_{T\to S}(T),S)) \end{aligned} \tag{3.24}$$

where λ_{idt} balances the importance of the identity loss term.

CycleGAN is improved to NICE-GAN [Chen et al., 2020a] where the encoder network of the generator is integrated into the discriminator. Thus, instead of computing another encoding step from the generated image, the discriminator directly inputs the latent code of the images. This not only lightens the architecture, but it also makes the training more stable since the backpropagation of the GAN loss is applied directly to the weights of the generator.

Another unsupervised GAN-related approach is proposed in [Fu et al., 2019] and aims at ensuring the robustness of the network to geometrical transformations of an image - *i.e.* rotations or vertical flipping. To this end, the Geometry-Consistent GAN (GCGAN) is based on a single generator-discriminator pair and on a novel geometry-consistency loss, L_{geom}. It consists in comparing the translation of a transformed image with the transformation of the translated image. Equation (3.25) gives the computation of the geometry-consistency loss by considering $f(.)$ as the geometric transformation and I as the ensemble of original images.

$$L_{geom}(I) = L_{MAE}(G(f(I)), f(G(I))) \tag{3.25}$$

Therefore, the training loss L_{GCG} of the generator of GCGAN is a linear combination of the GAN loss and the geometry-consistency loss as in Equation (3.26).

$$L_{GCG}(S,T) = L_{geom}(S) + \lambda L_{GAN}(G(S),T) \tag{3.26}$$

These various architectures have been widely studied and improved in computer vision to ensure better translation and style transfer capabilities. One critical challenge is to make the hallucination property beneficial for the translation task and to avoid the generation of artifacts which are unrelated to the actual input image. The authors of [Chen et al., 2022a] implement a CycleGAN-like architecture for synthetic-to-realistic image translation process but without using the cycle-consistency loss. Instead, the loss evaluates the style-content separation. First, it forces the low frequencies of the synthetic and realistic images to match because they are characteristic of the content of the image. Then, based on the discriminators, the

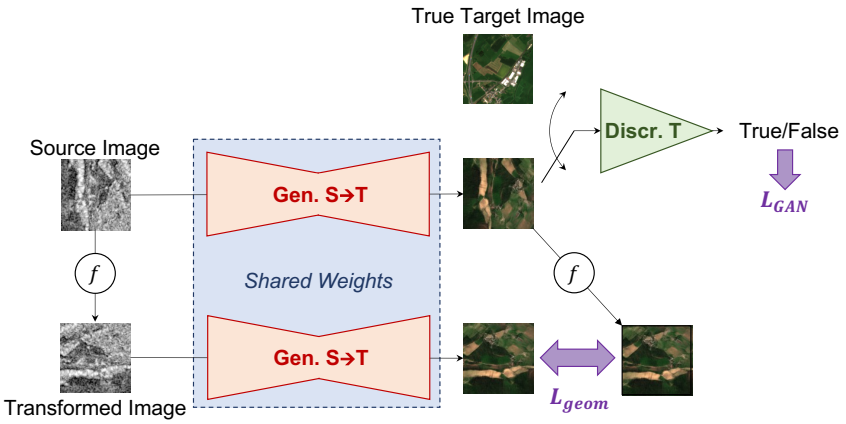


Figure 3.5: Adversarial training of GCGAN [Fu et al., 2019]

high frequencies are evaluated to match the real or synthetic distribution. Both synthetic-to-real image translation and style transfer are considered in [Zhang and Li, 2021] for dehazing photographs. Not only do the authors aim to increase the size of the real-world image dataset by generating pseudo-real images translated from synthetic images, but they also target the hazy-to-clean image transfer. To this end, the dual encoder-decoder network targets four tasks: hazy synthetic to clean synthetic image style transfer, hazy real to clean real image style transfer, hazy real to hazy real image autoencoding and hazy synthetic to hazy real image translation. The translation proposed in [Piva and Dubbelman, 2023] also addresses domain adaptation between synthetic and real images but applied to a downstream segmentation task. Several translation networks are trained to translate the synthetic images into multiple possible real images. These translations are then fed to a segmentation network trained with an ensemble learning strategy. A grayscale-to-RGB color translation is implemented in [Iizuka et al., 2016] by leveraging an aside classifier network. The latter aims at discriminating the classes in the input grayscale image based on the features extracted by the encoder of a style transfer network. The refined classification features are then given as input to the decoder of the translator to bring the semantic knowledge related to the composition of the scene and to reconstruct appropriate patterns with respect to it. Finally, the main contribution proposed in [Zhang et al., 2024] is a novel discriminator network. It inputs two overlapping patches from the input image and ensures that the feature extraction process in the overlapping region is similar. This approach enforces local consistency of the extraction process and thus better stability in the training.

These approaches have also been widely extended in remote sensing for many different tasks and especially for SAR-to-optical modality translation as reviewed in [Zhang et al., 2023b]. Despite the artifact generation issue, the modality translation approaches are mainly considered for image generation purpose but rarely for downstream tasks which explains that such irrelevant patterns are less considered. This is highlighted by the authors of [Zhao et al., 2022a] who experimented seven supervised and unsupervised neural networks for translating SAR images into optical ones. The conclusion of the study makes Pix2pix [Isola et al., 2017] and CycleGAN [Zhu et al., 2017] as the most credible, accurate and effective networks for the task. Therefore, in the following we propose a review of the supervised and unsupervised adversarial approaches applied to remote sensing images in the literature - mostly based on the two aforementioned architectures.

3.3.2 Supervised Adversarial Networks

Supervised adversarial modality translators, and in particular the Pix2pix [Isola et al., 2017] architecture, are used, studied and improved in remote sensing for their simple but effective generator (typically based on U-Net [Ronneberger et al., 2015]) and discriminator. The authors of [Bermudez et al., 2018] train this model for a SAR-to-optical translation which is further used for cloud removal. Similarly, this architecture is used in [Merkle et al., 2017] and [Merkle et al., 2018] for optical-to-SAR translation. The authors' goal

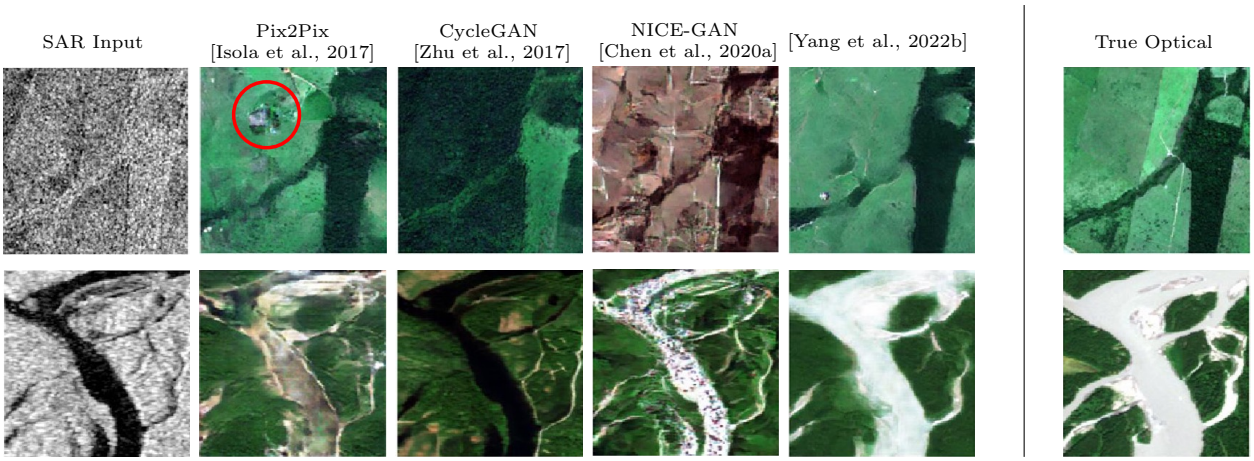


Figure 3.6: SAR-to-optical translation performed by four methods from the literature. Modified from [Yang et al., 2022b]. Red circled area illustrates a typical hallucinated artifact.

is to exploit the strong structural properties of SAR images to extract more salient keypoints than in the optical images and compute a more precise image registration. Nevertheless, since the translation between a SAR and an optical image is an ill-posed problem, the original U-Net architecture may not be effective enough for feature extraction.

Therefore, several authors propose to improve the generator of Pix2pix to allow the extraction of more complex features and the reconstruction of more relevant optical patterns. The authors of [Lee and Park, 2023] replace the U-Net architecture with a U-Net++ model [Zhou et al., 2018] which allows a better projection of the input modality features onto the output feature space. The SAR-to-optical translation is then used for military vehicles detection. The authors of [Darbaghshahi et al., 2022] propose to introduce dilated residual inception blocks in the latent space of the U-Net architecture to ensure a better projection of the SAR features. These blocks are composed of three parallel branches that apply convolutions with different dilation ratios to explore multiple receptive fields during the projection process. A second network is then applied to a cloudy optical image and its SAR-to-optical translation to fuse them and ensure accurate cloud removal. An optical style transfer is applied in [Fang et al., 2023] to perform change detection and to avoid domain adaptation issues due to the different solar exposition. The so-called Content-Invariant Translation network considers the images as a combination of content and style features. A content-only feature extraction is applied to the pre-event images followed by a style decoder network. The former is ensured by a content-attention loss which forces the features extracted from the pre- and post-event to be identical in unchanged regions. The style decoder is then trained to fool a standard discriminator. From another perspective, the proposals developed in [Yang et al., 2022b] are threefold: a two-branch generator, a multi-scale discriminator and a chromatic aberration loss. The former consists of a first high resolution branch with small convolutional kernels and a second downsampling based branch with residual blocks to extract low frequency features. The multi-scale discriminator ensures that the reconstruction resulting from the combination of the two branches matches the real optical features at local, medium and global scales with three different classifiers. Then the chromatic aberration loss term consists of blurring the true and generated optical images and comparing their low frequency patterns. Qualitative comparisons of this approach with three of the standard computer vision translators from the literature are available in Figure 3.6. This figure first illustrates one of the main issues with adversarial networks: the generation of hallucinated artifacts as red circled in the result of Pix2pix [Isola et al., 2017]. In addition, it shows the great performance of the approach proposed in [Yang et al., 2022b] even if the authors selected images with few high resolution patterns which limits the conclusions about the quality of the reconstructed textures and details in the image.

Regarding the last approach, several works propose to improve the performance of Pix2pix by modifying

the loss function during training. Indeed, while the authors of Pix2pix [Isola et al., 2017] state that the MAE loss leads to less blurred translations, the authors of [Li et al., 2020] propose to replace it by the Structural Similarity Index Measure (SSIM) loss. The latter is more suitable for the task of SAR-to-optical translation because it forces the translated image to be structurally consistent with the optical ground truth. Therefore, such an evaluation strengthens the extraction of the structural properties of the SAR images. For the same reasons, the authors of [Zhang et al., 2020b] propose to use a Discrete Fourier Transform (DFT)-based loss to preserve the structural properties of the SAR images in the grayscale optical reconstructions. In addition, they propose to apply a VGG19 [Simonyan and Zisserman, 2015] classifier network downstream of the translation task. This auxiliary network aims at discriminating five classes of landscapes and forces a consistent pattern generation with respect to them. Thus, the final loss is composed of the standard GAN loss, the DFT-based loss and the MAE loss computed on the features extracted from the translated image and its ground truth. The authors of [Wang et al., 2022a] also suggest to use an auxiliary network to compute the loss of the main translation network. Unlike [Zhang et al., 2020b], their auxiliary network is an optical autoencoder, where the decoder is used to guide that of the translator through feature comparison. In this way, the translator is forced to mimic the same feature reconstruction process as the optical autoencoder but based on a latent space extracted from SAR images. On another level, the loss proposed in [Mourchid et al., 2020] is based on two different discriminators. The first evaluates the high-resolution details of the image and is called the image discriminator - *i.e.* it acts like the standard GAN discriminators. The second takes as input a blurred version of the translated and true images. This so-called color discriminator is forced to focus only on the global features such as the colors and the brightness of the image.

As depicted by this review, a wide range of improvements are applied to the standard Pix2pix architecture not only from an architectural point of view but also from an optimization point of view by means of custom losses. Nevertheless, SAR-optical pairs are not always available due to sudden changes that can occur between their acquisitions (*e.g.* snowfall, natural disasters or field harvesting) which can limit, disturb and misguide the training of supervised translators.

3.3.3 Unsupervised Adversarial Networks

Unsupervised approaches have been developed to address the different issues associated with supervised learning but also to exploit an even wider range of data. The use of the CycleGAN architecture [Zhu et al., 2017] is widely explored to perform the dual translation process and leverage both the source and the target modality features during the training. The authors of [Du et al., 2022] even propose a semi-supervised approach that couples Pix2pix and CycleGAN architectures. This architectural combination allows to exploit the powerful training capabilities of the supervised network without being limited to images with reliable ground truth.

Although the authors of [Liu and Lei, 2018] propose to use directly the original architecture of CycleGAN to prove the possibility to train unsupervisely a network for the SAR-to-optical and optical-to-SAR translations, the main applications of CycleGAN in remote sensing consist of architectural modifications to improve its translation capabilities. In [Fuentes Reyes et al., 2019], the authors propose to include residual blocks from [He et al., 2016a] to increase the transfer of high resolution features and to facilitate the convergence of the network. Following the same objective, the authors of [Fu et al., 2021] introduce novel cascaded residual connections to enrich the reconstructions of the decoder. These connections consist, first, of inserting a skip connection between the input image and the pre-upsampling layers and, second, of inserting a standard skip connection between the encoder and the decoder after the upsampling stages. This alternation between the input and encoder skip connections allows the use of the input image at different decoding scales as a prior reminder while guiding it with the features extracted from the encoder. From a style transfer point of view, the authors of [Zhang et al., 2022] propose to include Style-Recalibration Modules (SRM) in the latent space of the CycleGAN generator and in its discriminator for winter-to-summer and reverse translation. The modules are used to separate content and style features. While the former are not modified by the architecture, the latter are progressively projected into the output style feature space. This is done by computing the statistical properties of the input style features and modifying them with learnable 1D convolutions. The modified features are then integrated with the content features to ensure the style projection.

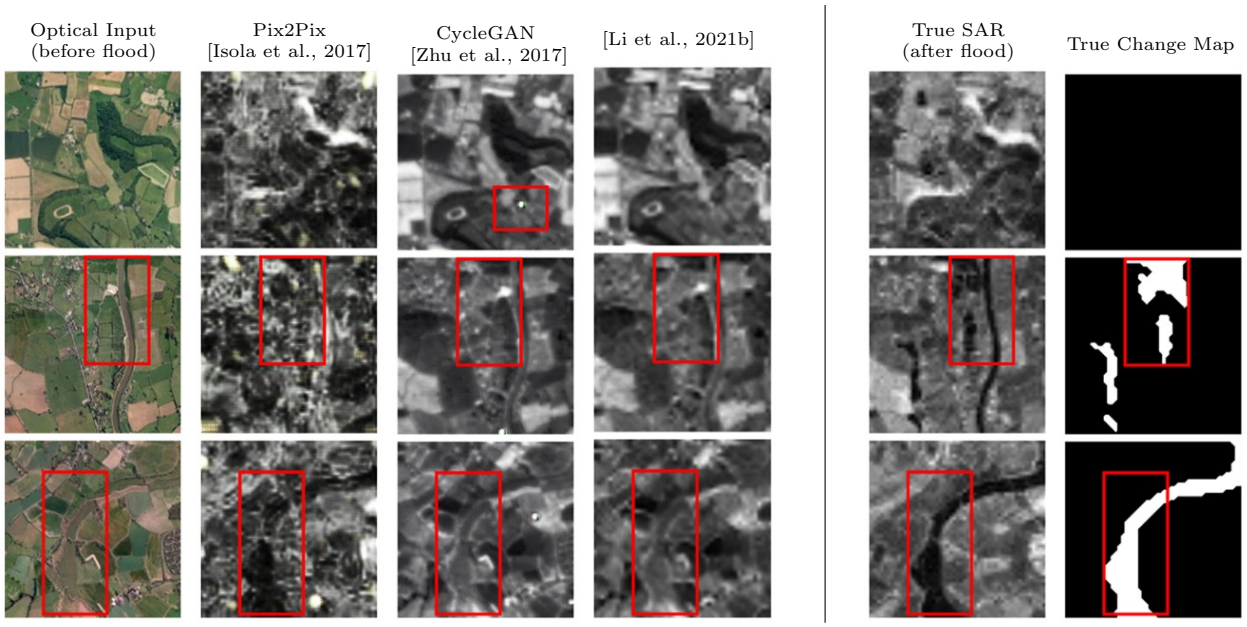


Figure 3.7: Optical-to-SAR translation performed by three methods from the literature intending to be used for downstream change detection. Modified from [Li et al., 2021b].

The second category of unsupervised neural networks for modality translation is based on NICE-GAN. The authors of [Luppino et al., 2021] propose to compare a non-adversarial CycleGAN and a NICE-GAN on their translation performance evaluated on a downstream change detection task. Similarly, the authors of [Zhao et al., 2023] propose a modality translation based on NICE-GAN. The specificity of the approach is the post-translation process referred to as Feature Refinement Module which aims to align the features of the translated image with those of a true image and to favor the confusion between the translation and the real domains. A final U-Net++ is applied to compute the change detection task. The same original NICE-GAN architecture is used for translation in [Li et al., 2021b]. Then, an improved version of U-Net++ is proposed by including separable convolutions [Chollet, 2017] and used to detect the changes between the pre-event translated image and the post-event true image for flood detection. Figure 3.7 gives the performance of the latter in terms of translation and in comparison to Pix2pix and CycleGAN. The annotation change map is also displayed to visually evaluate the distinguishability of the changes between the pre-event translated image and the post-event acquisition. Interestingly, all these NICE-GAN approaches are applied to an *a posteriori* change detection task. This is also the strategy which is explored in this thesis, in particular in Chapter 5, for the detection of slope instabilities despite cloudy optical images. More details on the issues related to the change detection task are given in Section 3.4.3.

Although adversarial approaches are still widely used today, they suffer from two main drawbacks that are particularly detrimental to downstream tasks: the instability of their training and the generation of hallucinated artifacts in the translated images. The latter is a significant issue since it creates patterns without any evidence for their actual presence. However, in order to make the output image as credible as possible and to fool the discriminator, the generator is highly encouraged to create such irrelevant patterns. This property has a strong impact on the further usability of the translated images. Since they contain inaccurate information, they cannot be entirely trusted which limits their use for downstream tasks and explains why, in this thesis, adversarial strategies are rapidly considered only as a contrast and credibility enhancement tool.

3.3.4 Towards Stable and Reliable Translations

Parallel to the rapid development of adversarial networks, many works intend to limit their drawbacks by completing or avoiding adversarial training.

A first approach is to incorporate signal and image processing techniques into the translation process. For instance, the wavelet decomposition of the images has already shown accurate performance for image-to-image modality translation in [Zhang and Xu, 2019]. Therefore, its multi-frequency aspect has been used in deep learning approaches to improve their performance and to guide generative adversarial networks. Based on this assumption, the authors of [Wei et al., 2023a] propose a dual branch network. The first branch is a standard Pix2pix translator with multi-scale feature combinations at the latent space level. The second branch inputs the wavelet decomposition of the SAR input image and filters the high frequency components to limit the impact of speckle noise on the translation and preserve only the structural properties of the SAR images. A decoder is then applied to the filtered components to produce an image used to enhance the structures in the output of the first branch. Thus, the wavelet decomposition guidance helps to limit the hallucinations of the network. Another approach proposed in [Li et al., 2022b] modifies the target of the modality translation. Instead of reconstructing the final optical image, the generator network is trained to reconstruct the wavelet decomposition of the optical image. In this way, the network is forced to focus on high-frequency patterns since they constitute the largest regions to be reconstructed by the network. In addition, to force the network to focus specifically on this task, they propose to reconstruct only the grayscale wavelet decomposition of the optical image and leave the coloration to a downstream style transfer network.

Besides networks based on wavelet decomposition, the use of transformer networks, and in particular Visual Transformers (ViT), are employed to regularize the reconstructions of adversarial networks. The authors of [Wang et al., 2022d] propose a generator that couples both local and global feature extractions from a CNN and a ViT branch, respectively. The ViT branch aims to avoid unreliable reconstructions with respect to the underlying land-cover. Therefore, the SAR features extracted by the ViT branch are classified by an adjacent shallow network to ensure relevant extractions. Then, the local and global features are exchanged between several layers of the CNN and ViT branches to limit the risk of artifact generation.

Another strategy is to develop multi-temporal approaches based on past paired images to better guide the reconstruction of adversarial networks. To this end, the authors of [He and Yokoya, 2018] replace the U-Net generator with a residual block CNN [He et al., 2016a] that takes as input a past SAR-optical pair together with the SAR image to be translated. The authors demonstrate a significant performance increase over the mono-temporal adversarial network. Past images are also exploited by the authors of [Liu et al., 2021a]. The images are fed to an external encoder that extracts attention maps to guide the reconstruction of the image by the main translator network. The authors also find that a non-adversarial training leads to an improvement in quantitative metrics despite less attractive visual results.

Based on the adversarial training issues and on the aforementioned experienced performance drop, several works are conducted to develop accurate modality translators trained without adversariality. The authors of [Luppino et al., 2024] propose to design a CycleGAN generator-only architecture. To enforce the alignment of the latent spaces and despite the lack of discriminators, they propose a novel code correlation loss term. The latter consists of first computing the distance between the affinity matrices of two input images from different modalities. Then, the loss forces the correlation between the latent vectors to match the calculated distance. For this purpose, the authors avoid the use of the discriminator in the architecture, in contrast to their work in [Luppino et al., 2021], but ensure modality confusion in the latent space. Another approach in [Pan et al., 2023] consists of a dual branch neural network for gradient and contrast feature extraction. The first one inputs the SAR image for structural feature extraction while the second one aims to mimic the optical extraction process by inputting a three-channel image resulting from the copy of the SAR image. An additional optical autoencoder is proposed to evaluate the features reconstructed by the decoder of the translator as in [Wang et al., 2022a]. The authors of [Feng et al., 2022] propose to combine both the super-resolution and the colorization of an optical input image. To this end, a style transfer based on a U-Net++ architecture is implemented. Its first peculiarity is the use of involution blocks [Li et al., 2021a] for the downsampling steps, which consist in determining the spatial convolution of each given pixel based on the evolution of its value along the channels. Then, sub-pixel convolutions [Shi et al., 2016] are employed for the upsampling steps and convolutional block attention module (CBAM) [Woo et al., 2018] for increasing the relevance of the features in the skip connections. The translation proposed in [Xia et al., 2024] consists in leveraging the spatial resolution of Sentinel-1 SAR images and the spectral resolution of MODIS multispectral images to generate Landsat-8 images with high spatial and spectral resolution. For

this purpose, both input modalities are processed in independent branches before being merged with an attention model. The decoder of the architecture, based on residual blocks [He et al., 2016a] and sub-pixel convolutions [Shi et al., 2016], allows the reconstruction of the final Landsat-8-like images. Finally, with the exponential development of diffusion models in computer vision [Sohl-Dickstein et al., 2015], which demonstrate better generation performance than GANs in [Dhariwal and Nichol, 2021], a recent diffusion model-based approach has emerged in [Bai et al., 2024]. The latter consists in conditioning the noisy map used for optical generation with a SAR image. This additional input data acts as the controllable vector in [Mirza and Osindero, 2014], but is suitable for translation and avoids adversarial training. The authors show promising performance in several landscapes which encourages for the development of diffusion models in remote sensing and in particular for the SAR-to-optical translation task.

3.3.5 Conclusion

As reviewed in this Section, the vast majority of SAR-to-optical translators, and more generally of modality translation or style transfer strategies, are implemented in an adversarial setting. Nevertheless, since these approaches are subject to a number of drawbacks (unstable training, artifact generation, drops in quantitative performance [Liu et al., 2021a]), novel approaches are emerging to regularize GAN-based networks or implement alternative strategies to ensure reliable generation despite a possible lack of credibility in the final images. Although image credibility can be of great interest in computer vision, the reconstructions obtained in remote sensing are expected to be reliable above all. Indeed, due to the downstream tasks applied to the translated images, the relevance of the reconstructed patterns and the absence of artifacts are of primary importance in remote sensing image translation. All these reasons justify that, in the Chapter 4 of this thesis, adversarial strategies are only used for contrast enhancement purpose but not for improvement in terms of patterns relevance. Thus, as soon as the relevance of the patterns are lowered due to an adversarial training for the sake of image credibility through the generation of hallucinated artifacts, the approach is considered as less interesting for the downstream targeted task: slope instability detection.

Some approaches even train the translator network to be consistent with a downstream task as in [Wang et al., 2022d] where the ViT features are classified by a subnetwork or in [Zhang et al., 2020b] where the final generated image is classified on the landscape to which it belongs. However, further studies should be conducted to ensure the relevance of the reconstructed patterns with the underlying land-cover, as it is known to affect the SAR-to-optical translations [Wang et al., 2022a]. Once such reliable translation algorithms exist, their use for further downstream tasks, and in particular for the identification of underlying classes, can be applied with more confidence.

3.4 Class Identification

In this Section, a description of the methods from the literature that deal with the detection and identification of classes in the input images is provided. This research area is of interest for the thesis for two main reasons. First, the use of an auxiliary classifier for a multitask learning setting is known to improve the performance of the main task, whether it is image generation [Dhariwal and Nichol, 2021, Odena et al., 2017], image captioning [Kandala et al., 2022] or modality translation and style transfer [Iizuka et al., 2016, Wang et al., 2022d, Zhang et al., 2020b]. Second, the identification of classes within images is directly linked to the final goal of this thesis - *i.e.* the detection of sudden and slow slope instabilities.

For the sake of clarity of the following discussion, and due to the variety of definitions that can be found in the literature, four types of class identification tasks are distinguished in this work and illustrated in Figure 3.8:

- **The classification** of an image corresponds to the labeling of the entire image depending on the presence of a class of interest in it (or classes for multi-label classification). Following the yellow path in Figure 3.8, the image is classified as containing or not landslides, without any further information about the spatial location of the susceptible landslides.

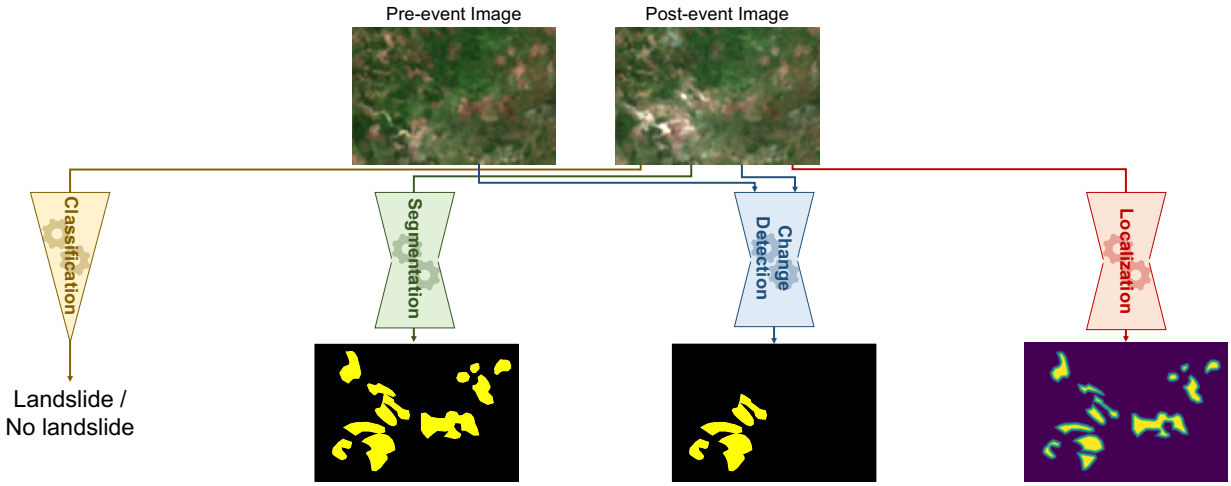


Figure 3.8: Scheme of the different types of class identification.

- **The segmentation** of an image refers to the process of separating an input data in several homogeneous regions. In the particular case of images, it consists of detouring the regions which differ from the rest of the image given a specific property of interest. The notion of classes only appears in the semantic segmentation process where each delineated region gets assigned a class. Following the green path in Figure 3.8, a binary segmentation is applied where the yellow regions belong to a landslide while the black ones are part of the background. By extension the concept of assigning a label to each independent pixel of the image is usually referred as segmentation. In the following, such strategies are considered as pixel classification to better distinguish segmentation maps generated from classification-based approaches from actual segmentation methods. The most important difference between segmentation maps and pixel classification maps is the consistency of the labels from a pixel to another which is not ensured by pixel classification approaches.

- **The change detection** task is the segmentation of the changes that occur between a pre-event image and a post-event image acquired over the same area of interest. Following the blue path in Figure 3.8, the landslides that were already present in the pre-event image are not segmented.

- **The localization** of a class in an image consists in highlighting the regions of the image that are susceptible to hold the given class. Thus, instead of a discrete (or *crisp*) label, each pixel is assigned a probability value of belonging to the class of interest, *i.e.* a *fuzzy* label. Following the red path in Figure 3.8, the boundaries of the landslides have smaller values than their centers highlighting the uncertainty of the detection in these regions.

The following discussion details the various related works of each of these class identification types except for localization which is the core of the following Section 3.5. Although these approaches are closely related, their respective peculiarities imply the implementation of different strategies.

3.4.1 Classification

The development of classification approaches is mainly used for pre-detection tasks. This image-wise labeling task is less time and memory consuming than segmentation and allows fast decisions. Therefore, it can be used as a pre-processing step before applying more complex pixel-wise approaches for local class identification.

To encourage for such deep learning classification strategies, the authors of [Sumbul et al., 2019] propose a novel dataset called BigEarthNet which consists of Sentinel-2 images and addresses multi-label land-cover classification. The relevance of the dataset is demonstrated by training an Inceptionv2 network [Szegedy et al., 2016] and a shallow CNN classifier. The dataset is updated in [Sumbul et al., 2021] to allow multimodal

land-cover classification from SAR and optical images. Both VGG [Simonyan and Zisserman, 2015] and ResNet [He et al., 2016a] classifiers are trained to demonstrate the usability of the dataset.

Nevertheless, at inference time, the availability of each modality can be a challenge, especially for optical images that may be missing due to cloud cover. Therefore, there is a growing interest in developing *hallucination* networks [Hoffman et al., 2016]. Far from being related to the issues of GANs, hallucination networks aim to extract features related to a missing modality but from another available modality. The authors of [Kumar et al., 2021] propose a four-step framework to train a land-cover classifier to extract multispectral features from a panchromatic image. First, two classifiers are trained separately and assigned to a single modality. Then, a conjoint training is performed to ensure the relevance of each extraction with respect to the other and to use both modalities for the classification task. In the third step, the multispectral modality is considered as missing and a third classifier is trained to mimic its features from a panchromatic image. This hallucination classifier is considered as a student network that should distill the knowledge from the teacher multispectral classifier. Finally, the hallucination and panchromatic classifiers are jointly tuned for the classification task. The same four-step training is applied in [Li et al., 2022d] for land-cover characterization based on multispectral and SAR images - considering the SAR image as missing at inference time. Their contribution lies in the distillation process of the third step where they extract the features from multiple layers in the classifiers and apply attention modules to better map the hallucinated features with the true missing features. The distillation of knowledge from the teacher to the student is also modified in [Wei et al., 2023b] as they explore the use of a loss comparing the distributions of the features rather than their actual values. This strategy is used for land-cover classification on hyperspectral and multispectral images.

Apart from hallucination networks, the authors of [Li et al., 2022a] focus on SAR images and propose a ResNet classifier based on the Shake-Shake strategy [Gastaldi, 2017] to detect the presence of vehicles on aerial images. The Shake-Shake architecture consists of ResNet blocks with two parallel convolutional branches. The two branches are linearly merged with a weight γ (resp. $1-\gamma$) for the first (resp. second) branch. The value of γ is randomly sampled at each forward pass which allows to increase the variability of the internal representations of the features. This strategy is used to train the network on an unbalanced dataset and to avoid bias and overfitting during the training. Further fine-tuning with a smaller balanced dataset is performed for the final classification task.

A two-stage training is also proposed in [Bountos et al., 2022a] to better detect volcanic deformations on InSAR images and provide early warning of the possibility of impending eruptions. To this end, the SimCLR framework [Chen et al., 2020b] is used to leverage contrastive learning and train a CNN feature extractor without the need for annotation. A simple fully connected layer is then trained on a few annotated samples based on the features extracted by the CNN. Volcanic unrest detection is also addressed in [Anantrasirichai et al., 2019] by training an AlexNet classifier. For data augmentation, the synthetic interferograms used for the training are wrapped with different gains and offsets. Thus, due to the rotation property of the phase, the images demonstrate multiple levels of fringes characterizing different possible ranges of ground displacements. Such a strategy aims at making the network robust to the phase rotation and to the variability of the target phenomena.

Natural disaster detection is also at the core of the study in [Ullo et al., 2019]. The authors propose a change detection-like approach based on a classification network. By considering a pre- and post-landslide collapse image from Sentinel-2, they train a 3D convolutional classifier to detect whether the input images contain recent landslides or not - as in the yellow path in Figure 3.8 but by inputting both images. The authors of [Zhang et al., 2018c] propose to train a Siamese classifier network [Bromley et al., 1993] based on digital surface models to detect whether an urban area contains changes or not. More details on Siamese networks are given in Section 3.4.3 and Figure 3.11. These digital surface models are derived from aerial images and 3D laser scanning of the area of interest. The latter two raw products are also used at inference time to compute the segmentation of the changed areas with surface-based growing algorithms.

Bitemporality is extended to multi-temporality in [Rambour et al., 2020] to detect floods in time series of SAR and optical images from a novel SEN12-FLOOD dataset. Both a ResNet50 [He et al., 2016a] and a ResNet50 coupled with Gated Recurrent Unit (GRU) [Cho et al., 2014] are trained to detect these floods and demonstrate the effectiveness of temporal networks. The authors of [Ienco et al., 2017] propose to apply a

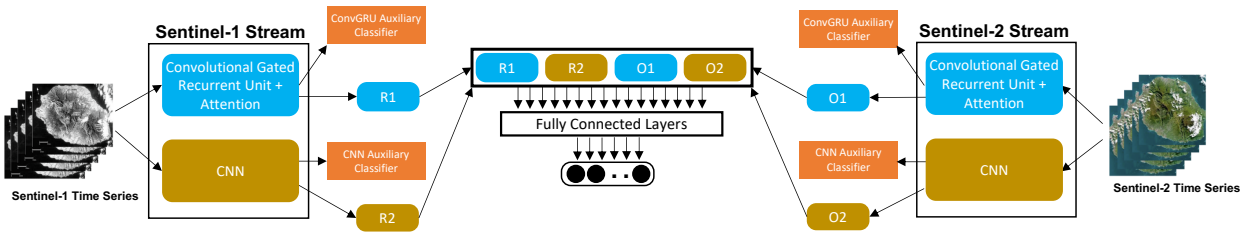


Figure 3.9: Scheme of the spatio-temporal approach proposed in [Ienco et al., 2019] for multimodal classification. Image modified from [Ienco et al., 2019].

standard instance segmentation approach as a pre-processing step and to use a classifier network to assign a land-cover label to each instance. To this end, two Long Short Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997] are trained on the mean and standard deviation of the instance pixel values for each image of the time series and on the temporal distribution of each instance pixel, respectively. A similar approach is proposed in [Mirmazloumi et al., 2022] to classify the type of ground displacement monitored by point scatterers from InSAR time series values. A CNN is applied to these time series to discriminate five classes: stable, linear, quadratic, bilinear or phase unwrapping error displacements. Ground deformation is also studied in [Brengman and Barnhart, 2021] with a CNN based on residual blocks. The objective of this classifier is to distinguish actual ground deformation patterns from atmospheric or topographic patterns in SAR interferograms. Transfer learning is used to transfer knowledge from synthetic to real data.

Although the previous approaches are typical classification approaches - *i.e.* assigning global labels to a given input - a wide range of methods leverage classifier networks for "segmentation" purposes and can be considered as pixel classification approaches. This strategy consists of applying a classifier network to each pixel and determining its class based on its temporal variations and/or spatial neighborhood. A sliding window strategy is then applied to ensure the classification of each pixel of the image. Yet, by nature, such a strategy cannot be considered as a segmentation since the consistency of the labels from a pixel to another is not guarantied with this approach. Therefore, the concept of *segment* or *region* is not integrated in the resulting pixel classification maps.

The authors of [Ndikumana et al., 2018] propose to extract the values of a given pixel in the VV and VH channels of a SAR time series. The vector is then fed into a series of LSTM [Hochreiter and Schmidhuber, 1997] to classify the underlying crop type. The same strategy is explored in [Kussul et al., 2017] to extract a temporal vector of Sentinel-1 values. They also propose to input the neighborhood of the given pixel from a Landsat-8 (optical) image to classify the underlying land-cover. Each input is fed into a 1D CNN and a 2D CNN to extract both temporal and spatial features, respectively. Land-cover characterization is also targeted by the authors of [Benedetti et al., 2018] who propose to extract the temporal vector from a Sentinel-2 time series and the spatial neighborhood from a SPOT6/7 image. The former is then input to a GRU-based branch [Cho et al., 2014] while the latter is input to a CNN branch. Instead of extracting only the temporal values in a single pixel, the authors of [Interdonato et al., 2018] use the whole time series of the neighborhood of a given pixel for the classification of the central pixel. The same GRU-CNN architecture is used to predict the underlying land-cover. The authors of [Ienco et al., 2019] propose the same strategy but leverage both SAR Sentinel-1 and optical Sentinel-2 time series. The modalities are processed independently in two separate two-branch networks as depicted in Figure 3.9. Each network is based on a CNN branch and a convolutional GRU branch. The latter consists of modules similar to the convolutional LSTM [Shi et al., 2015] but with GRU [Cho et al., 2014] units instead of LSTM units. The outputs of each branch from each modality are first classified by an independent classifier. They are also combined to be classified conjointly by a fifth classifier. The final decision results from the combination of the classifiers predictions.

From a purely spatial point of view, the authors of [Hu et al., 2022] propose to use convolutional LSTM [Shi et al., 2015] to classify hyperspectral images by replacing the time dimension with the spectral dimension of this modality. In addition, the authors propose to use pseudo-complex-valued layers because

they state that complex-networks have better generalization, easier optimization and faster learning characteristics. Similar convolutional networks are used in [Zhang et al., 2017] to deal with fully polarimetric SAR images and to classify crops despite the mono-temporal nature of the images. Finally, the authors of [Ghorbanzadeh et al., 2019] aim to detect the location of landslides using a classifier on both optical and topographic maps. Multiple window sizes are selected for the sliding window process to better detect long and thin landslides as well as large landslides.

As can be seen from the above approaches, the boundary between classification and segmentation is thin, since the segmentation can be seen as pixel-wise classification. Nevertheless, pixel classification approaches rely on classifier networks that are lighter, shallower and less resource demanding than pure segmentation networks.

3.4.2 Segmentation

Although pixel classification approaches show great advantages over segmentation networks, they still have some limitations. First, due to their shallower architecture, they cannot learn as many features as a segmentation network. Second, the sliding window strategy requires a forward classifier pass for each pixel of the input image which makes it more computationally expensive to generate large segmentation maps. Third, segmentation networks can be used to search features outside the immediate neighborhood of the pixel, allowing for non-local strategies. Finally, the lack of consistency of the labeling strategy from a pixel to another is critical since a slope instability is expected to be spatially continuous, which requires the concept of *regions* to be integrated in the mapping strategy. For all these reasons, standard segmentation networks that directly output the segmentation map are still widely used in the literature.

With the great development of annotated datasets based on standard land-cover maps (*e.g.* CORINE [Feranec et al., 2016], MODIS [Sulla-Menashe et al., 2019], 3DBAG [Dukai et al., 2019]), researchers increasingly consider the land-cover segmentation task as a toy example for validating novel architectures. The authors of [Wang et al., 2022b] propose to modify the standard encoder of the UNet++ architecture [Zhou et al., 2018] by exploring five different backbone encoders to obtain the crops segmentation map of optical images.

The land-cover segmentation from SAR images is targeted in [Ley et al., 2018]. To better address the SAR difficulties for the classification task, they propose a multi-task learning approach. They train both a SAR-to-optical translation and the land-cover segmentation from the same features extracted by the main encoder. The SAR-to-optical translation is used to force the encoder to extract optical-related features which are easier to classify than SAR-only features. Segmentation is then performed by a standard CNN decoder from the latent features. A multimodal strategy is also proposed in [Scheibenreif et al., 2022] where two vision transformers (ViT) are unsupervisely pre-trained to project a SAR and an optical image into the same latent feature space. Once the ViTs have converged, a decoder is trained on a limited number of annotated samples to segment the underlying land-cover. A similar approach is explored in [Montanaro et al., 2022] to project SAR and optical features into a common space and generate the segmentation map from a limited number of annotated samples. For this purpose, CNN networks are pre-trained with a contrastive learning strategy to ensure the alignment of latent vectors extracted from both modalities. Further pre-training is performed to ensure robustness against image degradation.

One of the major advantages of training multimodal approaches to align the latent spaces of each modality is related to the missing modality issue. The latter is addressed by the authors of [Kampffmeyer et al., 2018] by training an hallucination CNN network encoder [Hoffman et al., 2016] to extract optical features from a SAR image. In this way, the hallucination encoder can replace the optical encoder when the optical image is missing. Therefore, a land-cover segmentation map can be computed from a single SAR image by extracting SAR features in a SAR encoder and optical features from the hallucination encoder. The same issue is addressed in [Zheng et al., 2021] by training a prototype network. It consists of an encoder network trained to extract common features from two different modalities. Two additional encoders are trained with the same architecture to extract modality-specific features. Following these steps, the authors suggest to find the linear transformation of the weights of the prototype encoder to match the modality-specific weights. In this way, the missing modality specific features can be computed by extracting the common features with the prototype encoder. The linear transform is then performed to project them into the modality specific

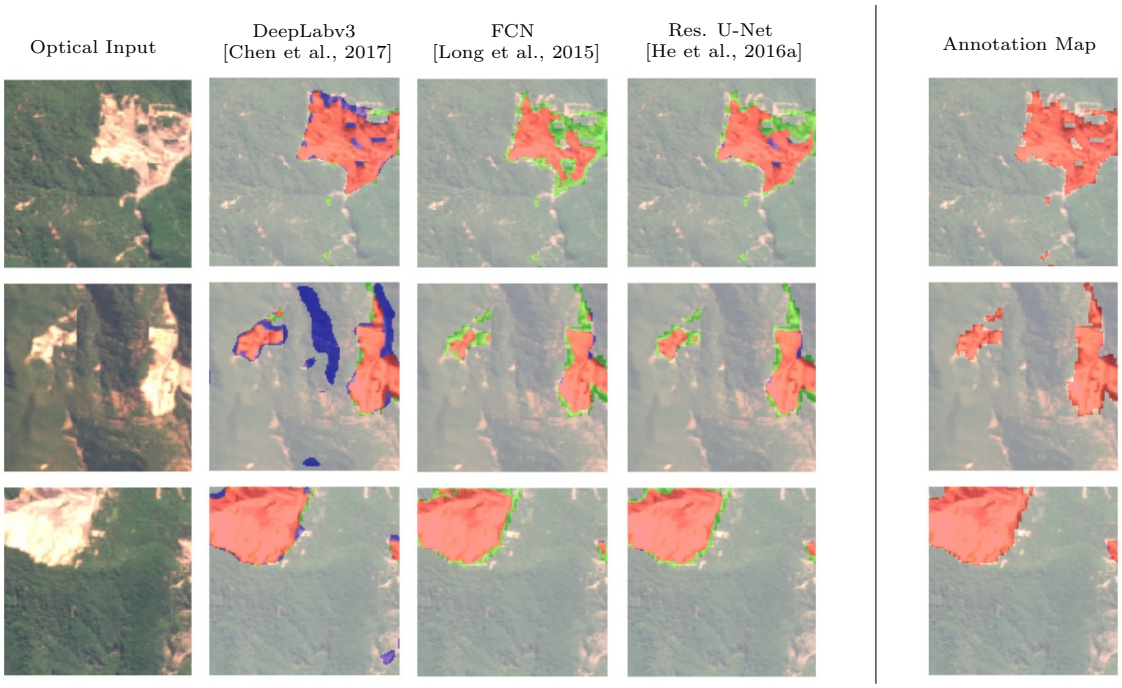


Figure 3.10: Landslide segmentations from standard networks from the literature. Image modified from [Ghorbanzadeh et al., 2022a]. Red, green and blue refer to true positives, false negatives and false positives, respectively.

feature space. The final segmentation map can be produced based on the actual features of the available modality and on the projected features.

Conversely, the authors of [Garnot et al., 2021] propose to use time series to extract features from both modalities. To this end, they propose a mid-fusion CNN which consists of extracting spatial features from each modality independently. The features are then concatenated according to their historical order and a temporal feature extraction is performed, resulting in the latent space of the network. The latter is finally used to generate a crop segmentation map with a CNN decoder.

In addition to the standard land-cover delineation, segmentation networks are particularly explored in the literature to identify and inventory the locations of natural phenomena at risk. The authors of [Bountos et al., 2022b] propose a new dataset, Hephaestus, for the identification of the unrest periods of 44 volcanoes on InSAR images and provide their segmentations. Both ViT and ResNet models are applied to the dataset as benchmark methods. The MMFlood dataset by the authors of [Montello et al., 2022] is created to address the segmentation of floods from Sentinel-1 SAR images, Digital Elevation Model (DEM) and hydrographic maps. Several networks are explored by combining four standard encoders and three standard decoders from the literature.

Landslide segmentation is addressed in [Xu et al., 2024] by creating the CAS dataset with manual annotations of post-collapse landslides. The latter is experimented to train standard segmentation networks such as Fully Convolutional Network (FCN) [Long et al., 2015] and DeepLabv3 [Chen et al., 2017]. The authors of [Ghorbanzadeh et al., 2022a] propose their Landslide4Sense dataset as a challenge to the community for novel contributions to the field with numerous baseline approaches. The qualitative performance of three of them can be found in Figure 3.10. As for the CAS dataset, it consists of the inventory of landslides from post-event optical images only. Four approaches are rewarded in [Ghorbanzadeh et al., 2022b] for completing the challenge. The first is a model ensemble approach that combines two encoder-decoder architectures based on a transformer and a CNN backbone, respectively, with a self-attention mechanism approach. The second is a two-stage training approach that ensures more relevant feature extraction in a U-Net architecture with

a transformer-based encoder. In the third approach, the authors propose a teacher-student framework to train the student network based on the pseudo-segmentation maps provided by the teacher network. Finally, the last method consists of a multispectral U-Net under a two-branch architecture. The first is dedicated to high resolution feature extraction while the second is based on the U-Net architecture and targets low resolution feature extraction.

Landslides are also inventoried in [Yang et al., 2022a] with a MaskRCNN network. The peculiarity of the training is the data augmentation which is based on the CutMix strategy [Yun et al., 2019] and consists in introducing landslide patches within clean images to increase the size of the dataset. Despite the lack of continuity between the landslide patch and the background image, the augmentation strategy allows a significant improvement in landslide segmentation. The authors of [Ghorbanzadeh et al., 2021] propose to investigate the performance of U-Net [Ronneberger et al., 2015] and Residual U-Net [He et al., 2016a] for landslide detection on optical images. They explore the effect of the size of the input image on the detection accuracy and thus the sensitivity of the networks to contextual information. In [Chen et al., 2024], the optical images are paired with SAR images and topographic maps (slope and aspect) to improve the accuracy of landslide detection. The authors propose to perform a first feature extraction step on the optical-SAR images separately from the topographic maps and to feed the concatenated features into a U-Net architecture. The specificity of the architecture is the use of an attention module in the skip connections to transfer more accurate multi-resolution features from the encoder to the decoder. Finally, the authors of [Zhao et al., 2022b] present a novel U-Net-like architecture with separable convolutions [Chollet, 2017] and Convolutional Block Attention Modules (CBAM) [Woo et al., 2018]. This architecture is implemented to detect deformation patterns (*e.g.* due to landslides or mines) in interferograms.

As depicted by the above reviews, both classification and segmentation approaches are involved in the detection and identification of natural phenomena. However, these approaches are implemented for static inventories, while in emergency cases the focus should be on the detection of novel phenomena only. Such selective segmentation can be achieved from bitemporal images with change detection approaches.

3.4.3 Change Detection

Similar to segmentation approaches, change detection methods can be addressed by using either classification or segmentation networks. Although change detection-like classifications [Ullo et al., 2019, Zhang et al., 2018c] are developed in Section 3.4.1, this Section focuses on the generation of change detection maps - *i.e.* to the pixel classification and segmentation approaches.

A novel framework is proposed by the authors of [Wang et al., 2019] to detect the changes in hyperspectral images from a pixel-wise approach. They first concatenate each input hyperspectral image with its abundance maps resulting from the unmixing process. Each pixel vector of the pre-event image is then compared to its corresponding pixel vector in the post-event image to compute an affinity matrix. The latter is finally fed to a CNN classifier responsible for detecting changes at the given pixel location. The authors of [Nava et al., 2021] propose a sliding window strategy to compare the classification performance of a CNN network depending on its input modality. Specifically, a first CNN is trained on a single RGB image and another on a pair of SAR images acquired before and after a landslide event. The authors find that despite the mono-temporal aspect of the optical training, the performance of the RGB CNN is still higher than that of the SAR change detection CNN.

A contrastive learning strategy is explored in [Zhao and Wang, 2023] to learn prototype semantic features of a pixel belonging to a changed or unchanged class. Bitemporal optical images are separately processed by a CNN and encoded by a transformer network. The two prototypes are then compared to the features and updated together with the parameters of the CNN to better separate each class. As experimented in [Chen and Bruzzone, 2022], the lack of homogeneity in the available modalities for the pre- and post-event images makes the change detection task even more challenging. Although this issue can be addressed with image modality translation approaches [Luppino et al., 2021, Li et al., 2021b, Luppino et al., 2024, Zhao et al., 2023] as described in Section 3.3, contrastive learning is an alternative that is explored in [Chen and Bruzzone, 2022] on homogeneous but also on heterogeneous inputs. Contrastive learning applied to heterogeneous data (SAR and optical in [Chen and Bruzzone, 2022]) aims to extract a common feature representation of the

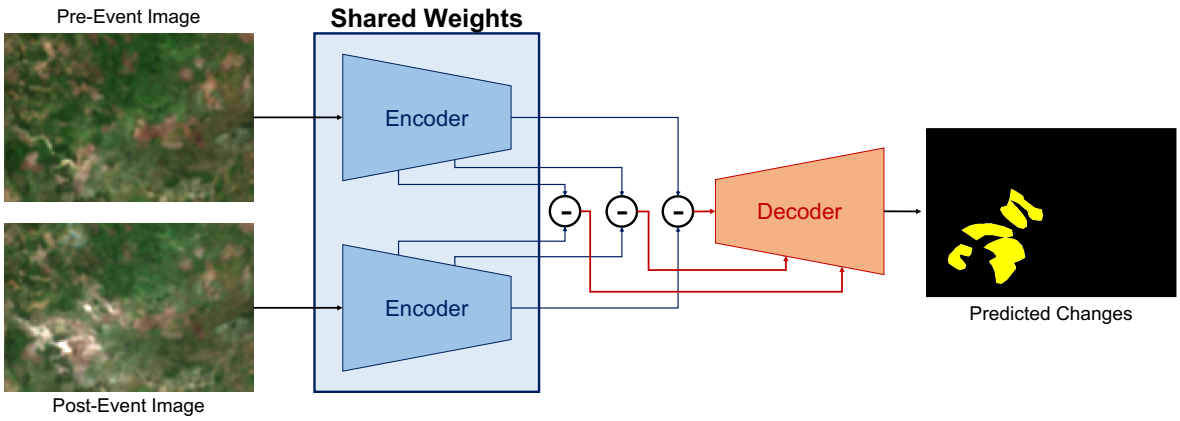


Figure 3.11: Typical Siamese encoder-decoder applied to the detection of novel landslides.

SAR and optical images by aligning the latent spaces of two separate encoders. The change map is then computed with a multi-size sliding window by computing the Euclidean distance between the latent vectors.

From a segmentation network point of view, standard approaches are based on Siamese networks [Bromley et al., 1993]. As illustrated in Figure 3.11, the latter consists of applying the same encoding to two different images and computing the difference of the extracted features to reveal their changes. In particular, coupled with a U-Net architecture [Ronneberger et al., 2015], the authors of [Ebel et al., 2021] propose the implementation of a multimodal Siamese network with two different encoders - one for the SAR images and another for the optical images. The SAR feature difference and the optical feature difference are then passed through the skip connections to the decoder to better map the land-cover changes. The authors of [Amankwah et al., 2022] explore two attention-based networks for detecting new landslides in Nepal. The first architecture is based on a Siamese U-Net++ architecture where attention modules are used to ensure a better combination of low- and high-level features in the decoder. The second is based on prior feature extraction and pyramidal attention module for multi-level feature extraction. The change map is derived from the Euclidean distance between the computed feature maps. A novel dataset for landslide detection on pairs of optical images is developed in [Nagendra et al., 2022] in several regions worldwide. To better address the problem of domain adaptation, the authors propose a learning without forgetting framework [Li and Hoiem, 2017] which consists of training a common ResNet encoder [He et al., 2016a] and multiple U-Net decoders dedicated to a single region. Each decoder is trained successively, ensuring that previously trained decoders maintain the same performance despite encoder updates through a knowledge distillation loss.

Other natural phenomena are also of interest for change detection such as floods and ice break events in [Ma et al., 2023] and forest fires in [Cambrin et al., 2023]. On the one hand, the authors of [Ma et al., 2023] propose to pre-classify the pixels based on an unsupervised clustering approach to train their network. The network is designed to extract spatial and frequency features from bitemporal SAR images and to use a fully connected layer to generate the change map. On the other hand, the authors of [Cambrin et al., 2023] propose a novel bitemporal optical dataset for wildfire delineation in California. Transformer and CNN approaches are used as baselines for this dataset for both segmentation and change detection tasks. The authors of [Gupta et al., 2019] focus on creating the xBD dataset to assess the damage caused by natural phenomena to buildings around the world including earthquakes, volcanic eruptions or tsunamis. Similarly, the authors of [Adriano et al., 2020] propose to evaluate the level of destruction of buildings affected by natural disasters. To this end, a U-Net architecture with two independent encoders is used to extract pre- and post-disaster features from time series. Five scenarios are explored in this work including monomodal, multimodal and crossmodal inputs selected between SAR and optical images. Finally, the authors of [Li et al., 2022c] propose a network to extract urban changes based on SAR, optical and Very High Resolution (VHR) optical images. They implement two U-Net++ [Zhou et al., 2018] architectures to separate the extraction of low resolution features in the SAR and optical images from the extraction of high resolution

features in the VHR image. The final change detection map is computed based on the multi-level features extracted by both networks.

The above discussion illustrates that despite the high similarity between the segmentation and change detection tasks, the challenges and the implemented approaches can be significantly different. Indeed, the bitemporal aspect of change detection requires more interactions between the inputs while multitemporal segmentation approaches focus mainly on monitoring the evolution of pixel values to better understand the temporal behavior of the underlying classes. Change detection is of particular interest for early response to triggered natural disasters which is why it is explored for sudden landslide detections in Chapter 5. However, neither segmentation nor change detection could be insightful enough for predicting future disasters, as it requires more physical meaning to be trusted.

3.4.4 Movement Detection

This Section focuses on the detection of movements or deformations that are precursors of natural disasters. Therefore, the privileged input information are the interferograms of the region of interest which can reveal very small deformations despite a coarse pixel resolution as mentioned in Section 2.4. Indeed, based on the physical properties of the SAR interferograms, deformation and velocity maps can be derived and used to monitor and predict future deformations and susceptible triggers of natural phenomena.

The authors of [Ma et al., 2020] propose to leverage a time series of interferograms to estimate the range of deformations of the Hong Kong International Airport. They implement a convolutional network based on residual blocks [He et al., 2016a] to extract the trend of deformation from 12 past interferograms and predict its future evolution for the 12 following time steps. InSAR is used in combination with further ground characteristics (*e.g.* hydrology, topography) in [Naghibi et al., 2022] to estimate the ground deformation based on extreme gradient boosting. A discrete approach is proposed in [Orlandi et al., 2022] to predict the velocity of ground deformations based on previous observations acquired by the Small Baseline Subset (SBAS) InSAR algorithm. Given a set of coordinates, a transformer network is trained to use all previous deformation values and output the future deformation for the given coordinates. The authors of [Shimosato and Ukita, 2021] propose to refine the InSAR derived land-subsidence images by using a CNN. This is necessary because InSAR images are affected by tropospheric delay which disturbs the interpretation of the final interferogram by introducing a variable phase offset. Therefore a CNN is trained to input the calculated land-subsidence and tropospheric delay maps to clean the former based on values measured at multiple *in situ* levelling points. A deformation score is calculated in [Valade et al., 2019] to estimate the state of activity of volcanoes. The authors train a ResNet autoencoder on synthetic data to input a wrapped interferogram and output the associated decorrelation mask and phase gradients. The objective is to use these outputs to unwrap the interferogram, compute the deformation map and obtain the corresponding score on real images.

Other approaches do not aim to compute the actual physical properties of the deformations but to localize them on the interferograms. The standard localization scheme is to output the probability map of a classifier/segmentation network resulting from the last softmax layer as illustrated by the red path in Figure 3.8. This strategy is explored in [Anantrasirichai et al., 2018] using a standard CNN classifier architecture for volcanic deformation detection on interferograms. These pixel-wise probabilities are converted into probability maps with a Gaussian merge of each patch. This approach is improved in [Anantrasirichai et al., 2019] to ensure better robustness against stratified and turbulent atmospheric patterns. In addition, the predictions resulting from real images are evaluated by volcanologist experts to reduce the domain shift between synthetic and real data and to incorporate expert knowledge into the network. These evaluations are further used to fine-tune the classifier on few real samples. Finally, the authors of [Biggs et al., 2021] apply the same strategy to predict the eruption of Turkish volcanoes. Due to the peculiarities of the terrain, the classifier network is also fine-tuned to be robust against snow cover and water vapor while estimating the deformations.

One of the main advantages of the localization approaches is their ability to map the actual deformation uncertainties. This property is actually important since the nearby pixels surrounding an area suffering from ground deformation are necessarily affected, even lightly, by the observed deformation. The uncertainty helps to map the extent of the susceptible danger while highlighting the main phenomenon of concern. Therefore,

this localization task is further explored in Chapter 6 for the detection of slow moving slope instabilities on interferograms. A standard computer vision approach to localization is neural network explainability with Class Activation Maps. Nonetheless, only a limited number of studies have been applied to remote sensing images to address the localization of suspicious areas affected by ground deformations.

3.5 Neural Networks Explainability

Since understanding the detections provided by neural networks is of particular interest in geoscience for trustworthiness, this section proposes to dive deeper into deep learning explainability methods. In this work, explainability is distinguished from interpretability. While the latter addresses the understanding of the inherent decision mechanism of deep networks, the explainability consists in finding more high-level clues to understand which components, regions or features of the input data guide the decision of the networks. Since this research area is extremely broad and in constant expansion, the choice is made to focus on a specific type of explainability approaches - namely, Class Activation Maps (CAMs). Due to their localization ability, their high performance and their applicability in numerous research areas, CAMs are considered as promising approaches for the task of slope instability localization addressed in this thesis. Therefore, the following discussions relate the existing state-of-the-art CAMs as well as their use for weakly supervised learning strategies. The Section concludes with the few approaches that address specific CAM-based constraints to ensure consistency between network decisions and their explanations.

3.5.1 Class Activation Maps

Class Activation Maps are specifically designed to understand the behavior of classifier networks with respect to their inputs. The main idea of a CAM is to highlight the most salient regions of the input image related to a given class of interest. Therefore, the generated heat maps indicate the regions that guide the network during the decision process. The pioneering CAM is developed in [Zhou et al., 2016]. Based on a standard classifier architecture, the authors suggest to use a Global Average Pooling (GAP) layer to convert the last convolutional features into a 1D vector. A final Fully Connected (FC) layer is used to classify the vector. Thus, focusing on the particular class of interest, the score for the class is computed by a linear combination of the components of the feature vector. The authors propose to apply the same linear combination to the final convolutional feature maps F_L and to generate the CAM of the given class as given in Equation (3.27).

$$CAM_L = \sum_{k=0}^{N_L} \alpha_L^k \odot F_L^k \tag{3.27}$$

where α_L^k is the linear weight associated with the kth feature map F_L^k of layer L. The main drawbacks of this CAM are its applicability only to the very last convolutional layer and its requirement of a GAP layer which implies a significant loss of information for the classification task. This approach is the starting point for the development of more sophisticated CAM methods which can be applied to different layers L but which can also associate the $\odot$ operation either with the standard multiplication between a scalar and a matrix or the Hadamard product between two matrices, denoted $\overset{\text{H}}{\odot}$ below. The main difference from one CAM to another lies in the calculation of the linear weights α^k. Three specific categories of CAMs can be identified.

3.5.1.1 Gradient-based CAMs

Gradient-based approaches are particularly inspired by the performance achieved by the authors of [Simonyan et al., 2014]. By computing the derivative of the class score with respect to each pixel of the input image, the authors are able to visualize each pixel contribution to the decision. With similar motivations, the authors of [Selvaraju et al., 2017] propose Grad-CAM. The Grad-CAM is one of the most widely used CAMs in the literature [Miao et al., 2023, Zhang and Ma, 2021, Qiao et al., 2023, Zhao et al., 2022b, Chen et al., 2024, Wargnier-Dauchelle et al., 2023]. Its computation consists of backpropagating the gradient of the class

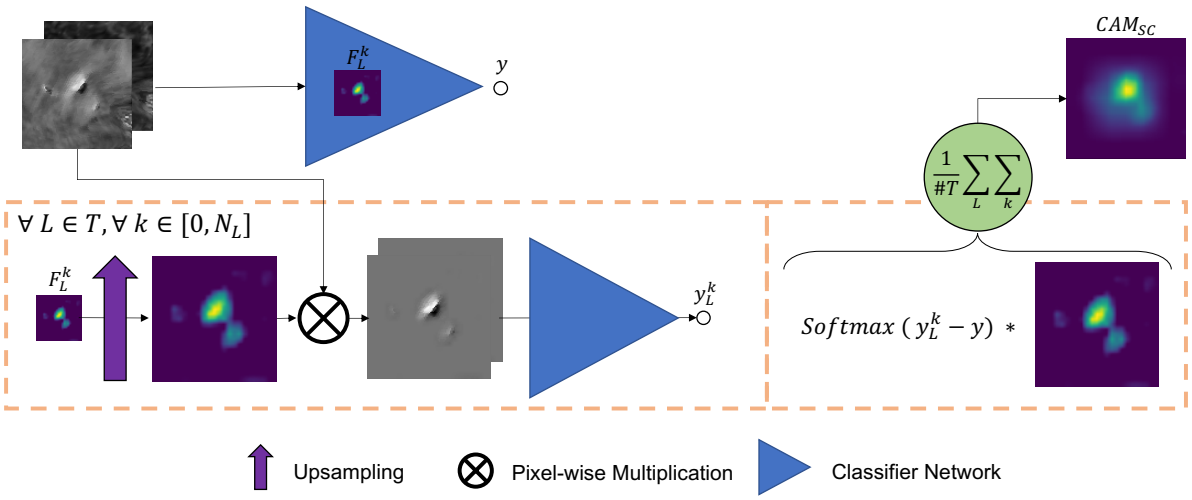


Figure 3.12: Generation of a multi-layer Score-CAM [Wang et al., 2020a] applied to the detection of moving areas on interferograms (see Section 6.4).

score to a given layer of interest. The weights associated with each feature map for the linear combination are then computed by averaging their respective gradient maps and assuming that positive gradients are associated with relevant features. Therefore, unlike the original CAM in [Zhou et al., 2016], it can be computed at any layer of any network and the Global Average Pooling layer is no longer required. As depicted by Equation (3.28), each linear weight α_{GC}^k is obtained by averaging the gradient map of the kth feature map.

$$\alpha_{GC}^k = \frac{1}{H \times W} \sum_{i=0}^{H} \sum_{j=0}^{W} \frac{\partial \hat{y}}{\partial F_{ij}^k} \tag{3.28}$$

where F_{ij}^k is the pixel of the kth feature map at location $(i, j) \in H \times W$ and $\hat{y}$ is the score output by the network for the given class of interest. The Grad-CAM is improved to Grad-CAM++ in [Chattopadhyay et al., 2018] by considering not only the first derivative but also the second derivative for calculating the weights. In this way, the authors allow the CAM to consider the size of the detected objects when computing the saliency map. However, despite the possibility to compute the Grad-CAM on all layers of the networks, the authors of [Rebuffi et al., 2020] demonstrate that it has a poor localization ability when applied in the early layers. The authors identify the averaging operation applied to the gradient maps as the cause of such a behavior, since it prevents from exploiting the high resolution of the features. Therefore, they suggest a more local approach by multiplying each feature map pixel by its corresponding gradient map pixel. Furthermore, in order to allow the CAM computation in any type of layer and not only in the convolutional ones, the authors propose to use the Frobenius norm $||.||_F$ to compute the final NormGrad. The overall CAM can thus be calculated as detailed in Equation (3.29).

$$NormGrad = ||\alpha_{NG}^k \overset{\text{H}}{\odot} F^k||_F \qquad \text{with} \qquad \alpha_{NG}^k = ReLU(\frac{\partial \hat{y}}{\partial F^k}) \tag{3.29}$$

Following similar motivations, the authors of [Jiang et al., 2021] propose the Layer-CAM which merges the weighted feature maps by the summation operation, *i.e.* using a close form of α_{NG}^k in Equation (3.27) and shifting $\odot$ to $\overset{\text{H}}{\odot}$.

3.5.1.2 Mask-based CAMs

Mask-based CAMs are implemented to avoid the problems related to gradients in deep learning such as the vanishing gradient issue which can reduce the reliability of the generated CAM in the early layers. A typical

example is implemented in [Wang et al., 2020a]. As depicted in Figure 3.12, the Score-CAM is generated by masking the input image $\mathcal{I}$ with each upsampled and normalized feature map $\tilde{F}^k$ of a target layer L. The masked images $\mathcal{I}^k = \mathcal{I} \overset{\text{H}}{\odot} \tilde{F}^k$ are then fed into the network to obtain the score $\hat{y}^k$ of the class of interest when a single feature region is input. The difference between each masked image score and the original score $\hat{y}$ reveals the importance of the feature map with respect to the class of interest. Therefore, the final Score-CAM is computed by linearly combining the feature maps weighted by the score difference, assuming that an increase in score indicates that the feature map contributes to the decision. To ensure only positive contributions, the softmax function is applied to the score vector before computing the linear weights α_{SC}^k as in Equation (3.30).

$$\alpha_{SC}^k = Softmax(\hat{y}^k - \hat{y}) \tag{3.30}$$

The authors of [Naidu et al., 2020] propose to improve the interactions between contextual features and the actual class of interest by progressively masking the image with the feature map. Thus, the Integrated Score-CAM (IS-CAM) masks the input image several times with weighted versions of the feature map - associating such a process with an integration over the feature map. In this way, the relevance of the feature maps can be better estimated to improve the quality and the accuracy of the final CAM. To avoid the computational explosion caused by the numerous forward passes proposed in [Naidu et al., 2020], the authors of [Wang et al., 2023] implement a positive-negative variant of Score-CAM. The input image is not only masked by the feature map to extract the features of interest but also by the opposite mask to extract only the rest of the image. The authors propose to compute the score of both masked images assuming that a feature map is of interest if the first mask increases the class score (positive impact) and the second mask decreases it (negative impact). From another perspective, the authors of [Desai and Ramaswamy, 2020] generate an Ablation-CAM. The latter consists in setting all activations of a given feature map in the target layer to 0. The decrease of the score obtained by the class of interest reveals the importance of the feature for the decision of the network. Thus, the decrease in class score experimented by each zeroed feature map is used to linearly combine the maps.

3.5.1.3 Complementarity CAMs

Unlike previous approaches, complementarity CAMs are based on two distinct classifiers. Their objective is to improve the localization ability of standard CAMs methods by searching for less salient but object-related regions. To this end, the authors of [Zhang et al., 2018b] propose the Adversarial Complementarity Learning (ACoL) framework. It consists of training a first classifier to extract the most salient region of the image. The second classifier is then trained on the same classification task by inputting the image where the salient regions highlighted by the CAM of the first classifier are masked. Therefore, the second classifier is forced to search for new salient regions to make an appropriate decision and thus focus on less salient but object-related features. The ReCAM proposed in [Chen et al., 2022b] assigns a different task to each classifier. While the first aims at multi-object classification with a final sigmoid layer, the second is trained to assign a single label per image with a final softmax layer. To guide the second classifier to focus on extracting finer regions, it is trained on the salient regions identified by the CAMs of the first classifier. For each detected class, the CAM is computed and used to generate a masked image. The second classifier is then trained to classify the masked images with the corresponding class. The authors of [Qin et al., 2022] adopt a different strategy based on the activation layers within the classifier architectures. Contrary to the ReLU activation functions used in their first classifier, the second classifier is based on Gaussian activation functions. The latter are used to remove both the low and high activations from the previous layers, keeping only the medium activations. In this way, the extracted features are still related to the target class but their saliency is lower. Therefore, the CAM of the second classifier can be used to complete the localization of the object obtained in the CAM generated from the first classifier.

Since the explainability of classifiers with CAM is still an ongoing research area, further approaches are developed by combining signal and image processing theories with neural networks. The authors of [Wang and Bunyak, 2023] suggest to use the Discrete Fourier Transform (DFT) to compute the DFT-CAM. Each feature map is first transformed to the frequency domain to compute an "orthogonality score" for each

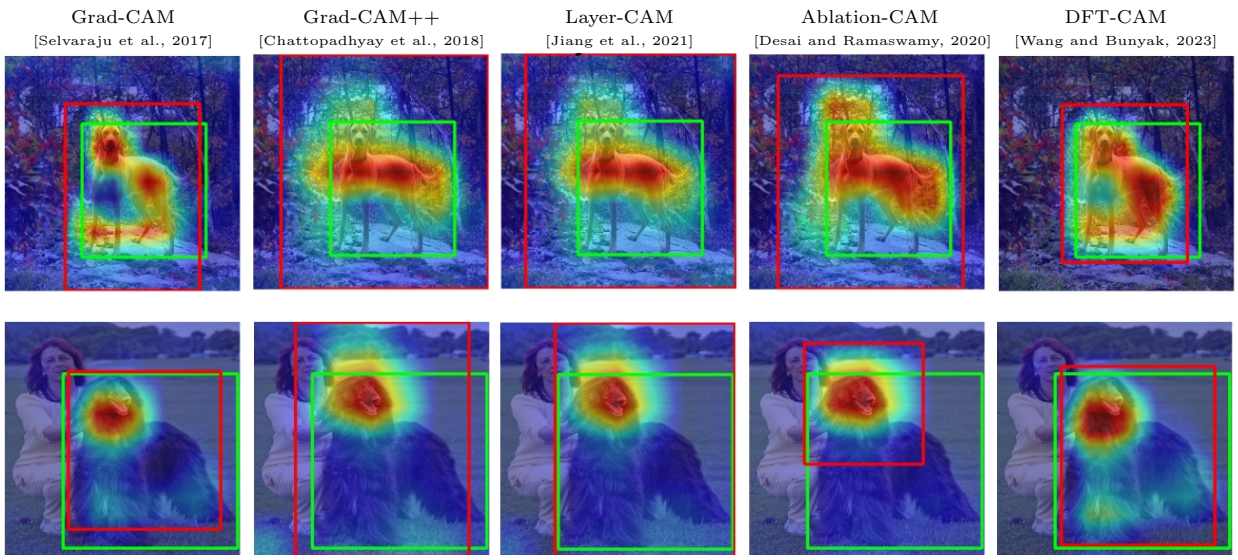


Figure 3.13: Performance of different CAMs on dogs localization. Modified from [Wang and Bunyak, 2023]. The green and red colors refer to the ground truth and predicted bounding boxes, respectively.

feature map. Contrary to the standard linear combination of all feature maps, the authors suggest to select only the most relevant features based on the orthogonality score and sum only these features to compute the CAM. The qualitative performance of DFT-CAM compared to gradient and mask-based methods is depicted in Figure 3.13 for dog localization. By their side, the authors of [Muhammad and Yeasin, 2020] propose the Eigen-CAM which computes the Singular Value Decomposition (SVD) of the activations of the last convolutional layer. The CAM is computed by projecting the activations onto the first eigenvector resulting from the SVD. The peculiarity of this approach compared to the other CAMs is that it is not class dependent. The projection on the first eigenvector only ensures the visualization of the most salient features of the image. Similarly, the authors propose to project the activations on the other eigenvectors to localize multiple objects on the same input image. The authors of [Chen and Sun, 2023] suggest a prototype based CAM. For a given layer of interest, multiple prototype vectors are kept in memory for each class. The computation of the cosine similarity between the feature map values of each pixel and the prototypes of the class of interest is used to generate the final LP-CAM.

Last but not least, some approaches propose to learn the weights used for the linear combination of the feature maps. Despite the fact that these approaches are closely related to attention mechanisms, they are still CAM-based methods. Opti-CAM is the typical example since the authors of [Zhang et al., 2023a] generates it by optimizing the weights of the linear combination of the feature maps during the training. As in the masking-based approach, the input image is first fed to the network to extract the features used for classification. Then, a 1D convolution is trained to merge all the feature maps and mask the input image. The masked image is then fed back to the network for classification. The backpropagation is used to learn the weights for the combination and compute the final Opti-CAM. The authors of [Lee et al., 2021] propose a more complex feature combination using an auxiliary network trained to explain the main classifier. The former inputs the features extracted by the classifier and computes the LFI-CAM which is used as a mask before the final classification of the main stream. Finally, the authors of [Zhu et al., 2023] propose to generate two different attention maps for object and background localization. Both maps are classified to ensure their separation and complementarity.

Despite the high performance of attention mechanisms in the literature, the resulting attention maps can be considered as learned maps. Since they are computed by backpropagation and optimized similarly to the standard neurons of the network, their explainability can be questioned. For this reason, attention mechanisms will not be considered in the following and the focus will be set on non-learned CAMs.

3.5.2 CAM-based Localization

All these CAMs are often evaluated on their ability to estimate the bounding boxes of the detected classes. Nevertheless, CAMs can be used for more precise localization tasks and are especially interesting for weakly supervised segmentations. Indeed, based on their localization ability, CAMs are used to generate pseudo-label maps which are considered as reference for training segmentation networks.

Typically, the authors of [Zhang et al., 2020a] propose to generate the original CAM [Zhou et al., 2016] of a classifier and to refine it with a dense Conditional Random Field (CRF) [Kraehenbuehl and Koltun, 2013] method. The resulting map is used to train a network for object segmentation in standard photographs. The UM-CAM framework proposed in [Fu et al., 2023] explores a similar method for the fetal brain segmentation in medical images of pregnant women. The peculiarity of the framework is the computation of two additional maps to distinguish the background from the target. For this purpose, seed pixels are selected as characteristic samples: the corners of the bounding box for the background and its centroid for the target. Each pixel is then compared to these seeds to generate the two additional maps. Finally, the segmentation network is trained to match both the CAM and the seed-based maps.

From a remote sensing point of view, pseudo-labeling is a key tool to address the lack of annotation maps in the datasets while leveraging their image-level labels as in BigEarthNet and BigEarthNet-MM [Sumbul et al., 2019, Sumbul et al., 2021]. For example, the authors of [Miao et al., 2023] propose a teacher-student approach for land-cover mapping. While the student network is optimized by a standard stochastic gradient descent, the teacher network derives its weights from the student network based on the exponential moving average. This strategy is explored to generate pseudo-label maps with the Grad-CAM of the teacher network and to semi-supervisely train the student network with labeled and pseudo-labeled samples.

The mapping of urban areas is treated in [Liu and Zhang, 2023]. As for standard pseudo-labeling methods, they train a first classifier to generate a multi-layer CAM based on features extracted in different layers. This pseudo-annotation map is used to train a segmentation network designed as a dual-stream architecture. The classifier and both streams are trained based on an interactive loss to ensure consistency between the predictions of the classifier and the generated segmentation. The dual-stream segmentation network consists of a standard decoder architecture branch and a branch built on a series of feedback blocks [Li et al., 2019]. The latter branch is borrowed from [Zhang and Ma, 2021] where the authors propose to train a feedback block-based network for the detection of urban areas on optical images. They use the Grad-CAM of an aside classifier in two ways. First, it is considered for data augmentation by pseudo-labeling unlabeled images. Second, the authors propose a curriculum learning strategy [Bengio et al., 2009] which consists in training the network on easy samples for the first epochs and include progressively harder samples from one epoch to another. In [Zhang and Ma, 2021], the authors propose to significantly noise the pseudo-label maps so that the network focuses on a coarse localization of urban areas. The noise is progressively removed during the training to compute more accurate segmentation boundaries. A curriculum learning strategy is also proposed in [Liu and Zhang, 2022]. The authors propose to modify the threshold used to binarize the pseudo-label CAM during the training of the segmentation network. The generated ACoL-like CAM [Zhang et al., 2018b] ranges from 0 to 1 so that pixels identified near 0.5 are uncertain pixels. To account for this uncertainty during the training, the authors propose not to label the uncertain regions based on a threshold that is progressively reduced during the training. This strategy forces the network to first learn the features of the regions with the highest/lowest probability of being urban areas and to progressively refine the segmentation by including uncertain pixels.

Finally, the authors of [Qiao et al., 2023] even propose to consider the Grad-CAM as an actual segmentation approach. They trained a classifier network to detect damaged buildings after an earthquake event. Based on a multiscale feature extraction and a spatial correlation step ensured by attention mechanisms, a fully connected layer is used for classification. Once the classification identifies the damaged building class, the Grad-CAM of the last convolutional layer is computed and refined with a dense Conditional Random Field approach [Kraehenbuehl and Koltun, 2013] to obtain the final segmentation map.

Nevertheless, although the pseudo-labeling strategy is studied for weakly supervised learning, CAMs are used in geosciences to give consistency to the proposed approaches. Indeed, the detection of ground

deformations by a network should provide more clues that can be trusted by experts. Therefore, some authors propose to ensure that the detection of their algorithms is right for the right reasons by visualizing the CAM in [Brengman and Barnhart, 2021] or the Grad-CAM in [Zhao et al., 2022b, Chen et al., 2024].

3.5.3 Right for the Right Reasons

To the best of our knowledge and despite the need for explanation in the geosciences and more generally in remote sensing research, the use of CAMs to train a neural network has not been explored except for pseudo-labeling purposes. For this reason, the following paragraphs will focus on approaches developed in computer vision. Some works have already intended to use the CAM as part of the loss to guide the network but also to constrain the CAM to match prior expectations.

From a guidance point of view, the contributions often consist in comparing the CAMs of two networks to ensure a better knowledge transfer, a more generalizable performance or an improved complementarity and consistency of the networks. The authors of Grad-CAM++ [Chattopadhyay et al., 2018] propose to apply their CAM to a teacher-student setting. The training of the student network is performed by ensuring that its Grad-CAM++ matches that of the teacher network to improve knowledge transfer capabilities. The training of the classifiers used to generate the complementarity CAM in [Qin et al., 2022] is also based on comparing their CAMs to ensure their complementarity. The same comparison strategy is employed in [Jo and Yu, 2021] for the Puzzle-CAM framework. The latter consists in dividing an input image into four non-overlapping tiles. Each tile is fed to the classifier network to generate its respective CAM. The authors propose to compare the CAM of the whole image with the CAM reconstructed from the four tiles. In this way, the objective is to ensure the extraction of more consistent features in the classifier network. The SEAM framework proposed in [Wang et al., 2020b] aims to increase the robustness of a Siamese network to image transformations. To this end, they propose to input both an original and affine transformed image to the classifier. The authors also implement an attention mechanism to enrich the CAM generated from the classifier with more details. Both the CNN and the attention module are trained by comparing the raw CAMs and the detail-enhanced CAMs of the original and transformed images. The authors of [Morales et al., 2021] propose to use the CAM as a data augmentation strategy. They suggest feeding the network with both standard images and images in which the salient regions highlighted by the CAM have been erased. This strategy forces the network to look for more specific features and increases the relevance of the final generated CAM. To improve the generalizability of the network, the samples are randomly selected from one epoch to another. Thus, each image is randomly fed to the network in its original or saliency-erased form.

The authors of [Ross et al., 2017] pioneer the development of explainability constraint approaches. Using several toy examples, they demonstrate that training the network to match an expected annotation map with the explanation provided by the gradients of the classification score with respect to the input image [Simonyan et al., 2014] strengthens the performance of the network and makes it more trustworthy. From an unsupervised point of view, the authors of [Erion et al., 2021] propose to constrain the results of the expected gradient method. Three attribution priors are described depending on the input data: pixel, graph or sparse attribution priors. In particular, the pixel attribution prior ensures that the expected gradients of a pixel are similar to its neighbors to ensure their consistency. Similarly, the authors of [Schöttl, 2020] propose three novel loss terms to constrain the CAM generated by a network despite the lack of annotation maps. Thus, loss terms based on the entropy of the CAM, its ellipsoidal area or its dispersion are explored for the training. The authors found that entropy is the most important property since training the network with it alone contributes to reducing the other two loss terms. From another perspective, a Grad-CAM constraint is proposed in [Wargnier-Dauchelle et al., 2023] to detect medical images showing brain tumors and localize the disease. Due to the lack of annotation maps, the constraint is applied only to non-pathological samples and consists in ensuring that the Grad-CAM generated from such samples leads to zero maps. In this way, the network is forced to extract features that are characteristic of tumors and avoid healthy ones. Figure 3.14 illustrates the tumor localization performance of explainability constrained networks compared to the Grad-CAM performed *a posteriori* of the training. Finally, the authors of [Boyd et al., 2023] propose a fully supervised training to detect whether an input image is real or generated by a GAN. To this end, they ask a number of participants to perform the same task and to highlight the regions of the images

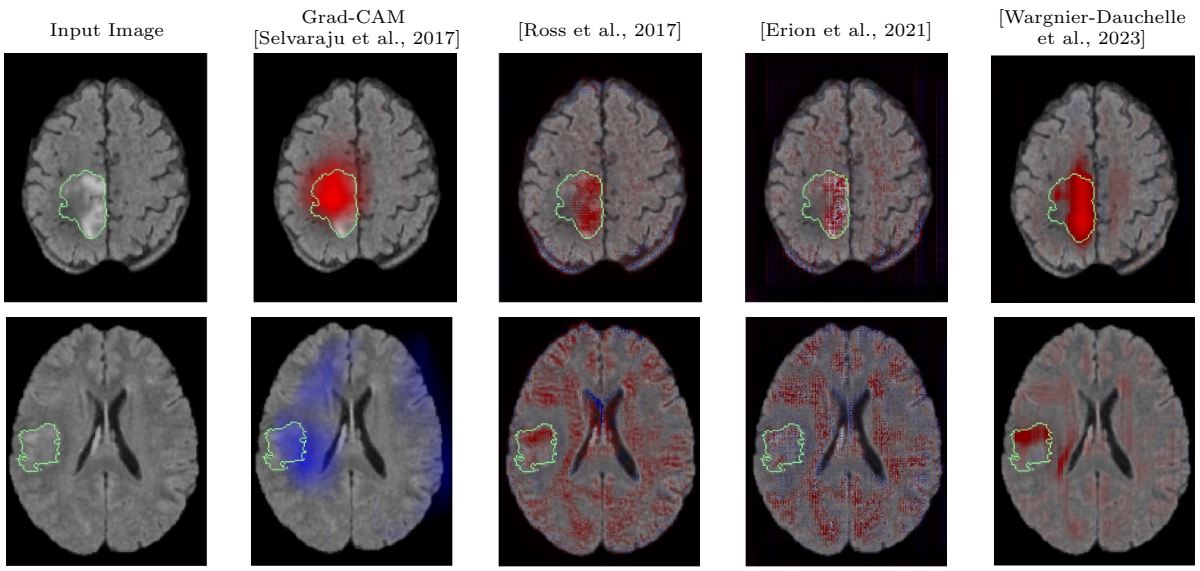


Figure 3.14: Performance of networks trained with an explainability constraint - except for Grad-CAM. Modified from [Wargnier-Dauchelle et al., 2023]. Colors range from blue to red which highlights healthy and pathological localization, respectively. The green delineation represents the ground truth tumor.

that support their decisions. These so-called human-saliency maps are finally used to constrain the CAM generated by the classifier network during its training.

3.6 Conclusion

In this Chapter, the focus was set on detailing the deep learning approaches from the literature that are related to the contributions proposed in this thesis. The state of the art in modality translation, class identification and neural network explainability is reviewed in order to better evaluate the existing works and their respective drawbacks. In particular, modality translation based on Generative Adversarial Networks (GAN) has shown outstanding performance in generating credible optical images. The counterpart of this strategy is the creation of unreliable artifacts which can significantly affect the understanding of the scene and therefore threaten the further class identification task. The latter is addressed in many ways in the literature depending on the targeted sub-task. The most common method is based on the use of residual blocks to reduce vanishing gradient issues and to ensure a faster convergence of the approach, *e.g.* in the Siamese networks used for change detection. Many approaches also use attention mechanisms and transformers which are heavier networks than standard CNN, even for simple classification tasks. Moreover, neither standard nor heavy networks are easy to understand for non-experts in the field which leads to many discussions and questions about the reliability of the methods. Therefore, explainability methods and especially Class Activation Maps (CAM) are proposed to better understand the decisions made by a network by highlighting the regions that support them. However, explaining the decisions *a posteriori* only verifies but does not ensure that the network is actually trained to extract relevant features. To this end, some approaches have been proposed to constrain the explainability of neural networks in computer vision but they have not been extended to geosciences or remote sensing research areas. Finally, the relevance of the explainability approaches depends mainly on the quality of the classification method which in turn is affected by the performance of the modality translation provided by the translator network. Therefore, an accurate translation is the keystone to ensure the reliability and accuracy of the downstream tasks for the monitoring of slope instability with multimodal images and exploit the acquisition of SAR images appropriately when no optical data is available.

Chapter 4

SAR-to-Optical Modality Translation

4.1 Introduction

Before addressing slope instability detection, a special attention should be paid to the weather sensitivity of optical imagery. Indeed, slope instability collapses usually occur during heavy rainfall events which induce large cloud cover and therefore a poor visibility on optical images. To address this issue, the first contribution of this thesis is the development of a novel SAR-to-optical translator neural network. Yet, slope instabilities are usually located on steep slopes in mountainous regions which cause strong distortions in SAR images. Such a constraint is rarely considered in the standard SAR-to-optical translation methods from the literature. Therefore, the proposals presented in this chapter aim not only at SAR-to-optical translation but they also aim at overcoming distortion effects and reconstructing optical images without unreliable artifacts. Importantly, since the translations are intended to be used for slope instability detection, the focus is especially set on reconstructing relevant and reliable patterns with less emphasis on their visual attractiveness and credibility. Multiple metrics are employed to measure the relevance of the performance from a radiometrical point of view (Mean Absolute Error (MAE) and Root/Mean Square Error (R/MSE)), from a structural point of view (Structural Similarity Index Measure with different window sizes (SSIM-x)) and from an image distortion point of view (Peak Signal-to-Noise Ratio (PSNR)). The Fréchet Inception Distance (FID) is the only metric implemented for measuring the credibility of the images based on the Inception network. Yet, this metric is considered as a secondary metric since the others highlight the actual relevance of the reconstructed patterns.

The chapter first introduces the different datasets of interest in Section 4.2 before describing the backbone architecture of the proposed translator in Section 4.3. Architectural modifications to improve the relevance of the translations are proposed in Section 4.4. Finally, the credibility and contrast enhancement strategies are detailed in Section 4.5.

4.2 Materials

To ensure the reliability and perspicacity of the proposed neural networks, three different datasets are used for training. This Section provides a description of the latter and introduces their respective interest for the study.

4.2.1 SpaceNet 6 - Urban Distorted Context

The SpaceNet 6 dataset described in [Shermeyer et al., 2020] was originally released for the sake of a challenge. The latter consists in leveraging a multi-sensor dataset to extract building segmentations over Rotterdam, the Netherlands. The dataset available on Kaggle[1] is commonly used in the literature [Zheng et al., 2021].

[1] https://www.kaggle.com/sandhiwangiyana/spacenet-6-multisensor-allweather-mapping, last visited on June 28, 2024.

SpaceNet 6 consists of 3,338 pairs of 450×900 pixel patches of SAR and optical modalities. The SAR images were acquired by Capella Space and Metasensing with a north- and south-facing airborne sensor on three different dates: August 4th, 23rd and 24th, 2019. The fully polarimetric sensor on board allowed for the acquisition of all co- and cross-polarizations (HH, VV and HV, VH) with a half-meter spatial resolution in the X-band. To limit the changes that can occur between modality acquisition dates, optical images were acquired by the Maxar Worldview-2 satellite on the 31st of August 2019. They are composed of four pan-sharpened bands - Red, Green, Blue and Near InfraRed (RGBNIR). Pan-sharpening is applied to ensure that the optical images match the half-meter resolution of the SAR images. Both modalities were coregistered and cropped to produce the multimodal pairs. Each pair is associated to an annotation map of the buildings obtained from the 3DBAG [Dukai et al., 2019] land-cover map.

The main interest of this dataset for the current study is the presence of strong SAR distortions in the images. Indeed, Figure 2.10 in Section 2.3.2.2 illustrates the case of a SpaceNet 6 image taken over vertical cylindrical tanks. The latter produce strong layover and shadow distortions represented by the bright circles and large dark regions in the SAR image. The images suffering from such artifacts have been isolated in a sub-dataset available in IEEE Dataport[2]. The latter is composed of 333 image pairs where the SAR image channels are clipped based on their respective mean μ_i and standard deviation σ_i in the range $[\mu_i - 3\sigma_i, \mu_i + 3\sigma_i]$ to reduce their strong dynamics. Each image of the pair is then normalized between 0 and 1.

This dataset is used to evaluate the distortion robustness of the proposed SAR-to-optical translator in an urban area. Since the latter show geometric patterns, it avoids additional constraints that can be observed in rural areas (*e.g.* the seasonal variations or the wide textural and structural varieties) which helps to focus the study only on the distortion robustness assessment.

4.2.2 BigEarthNet-MM - Rural and Annotated Dataset

Typical SAR-to-optical translation neural networks are applied to very large datasets but without any further context or annotation than the image pairing (*e.g.* SEN12MS [Schmitt et al., 2019] in [Montanaro et al., 2022, Scheibenreif et al., 2022, Jain et al., 2021]). Nevertheless, the translation task is often used as a preprocessing step for further downstream tasks. Therefore, the BigEarthNet-MM dataset [Sumbul et al., 2021] is selected in this work both for its size which limits the risk of overfitting and for its ability to evaluate the performance of the network from a downstream task perspective. The dataset is provided by the authors as an update of the BigEarthNet dataset [Sumbul et al., 2019] which was aimed at land-cover recognition from optical images only. The dataset gets updated not only to allow for all-weather classification but also to encourage for multimodal approaches in the field. Both datasets are available for download from the BigEarthNet website[3].

The BigEarthNet-MM dataset consists of more than 590,000 pairs of SAR and multispectral images recorded by the Sentinel-1 and Sentinel-2 satellites. The 120×120px patches were acquired from June 2017 to May 2018 over 10 different European countries. In addition, in order to ensure relevant correspondence and to avoid abrupt changes in between SAR and optical acquisitions, the images have been selected as temporally close as possible. The SAR images are Ground Range Detected (GRD) products in Interferometric Wide (IW) swath mode provided by the ESA. The logarithmic transform of both the VV and VH channels are provided in the dataset. On the other hand, the twelve channels of the Level-2A products of Sentinel-2 are available for use. Each image pair is annotated by using the CORINE 2018 land-cover dataset [Feranec et al., 2016] which contains 44 classes. However, due to the resolution of Sentinel images (10 meters) and due to the mono-date characteristic of the dataset, some classes cannot be properly distinguished from the others. Therefore, the authors have merged some classes together in order to propose a simplified land-cover consisting of 19 labels.

For memory and computational time reasons, a subset of 195,109 images is randomly extracted from the original dataset. In particular, since the SAR-to-optical translation is of primary interest here, all pairs with cloudy and snowy patterns are excluded from the subset to avoid unreliable optical ground truth. This

[2]https://ieee-dataport.org/documents/can-artificial-intelligence-untangle-distorted-and-compressed-geometries-associated-sar, last visited on June 28, 2024.

[3]https://bigearth.net/, last visited on June 28, 2024.



(a) Sentinel-1 SAR image (b) SRTM DEM image (c) Sentinel-2 optical image

Figure 4.1: Sample extracted from the Mountain Dataset [Bralet et al., 2024c]. The SAR image is colored with [0, VV, VH] as RGB channels.

dataset is first used in this chapter to verify the applicability of the novel approach to a more complex context - rural areas. Indeed, urban areas have deterministic and geometric patterns: straight streets, rectangular buildings or circular tanks. But, in a rural area, the context is more variable and less structured which is similar to that of mountains where slope instabilities are likely to occur. Then, Chapter 5 proposes a further use of the dataset to exploit the availability of the annotations for both the training and the evaluation of the translator.

4.2.3 Mountain Dataset - Typical Slope Instability Regions

Finally, we create a novel dataset to evaluate the performance of the SAR-to-optical translators in mountainous regions. The dataset contains triplets of SAR, Digital Elevation Model (DEM) and optical patches. The DEM is considered in this dataset to provide more information to the network about the susceptible locations of geometrical distortions, as it is directly correlated to the underlying slopes. Since the DEM is required for the orthorectification of the SAR images, it is a minor additional constraint at the time of inference. The proposed mountain dataset is freely available on IEEE Dataport[4] and first presented in [Bralet et al., 2024c].

The dataset is extracted from 36 triplets of images acquired over the northern French Alps between 2018 and 2021. The selected DEM is obtained from the Shuttle Radar Topography Mission (SRTM) [Farr et al., 2007] which aims to map elevation over a large part of the Earth. The SRTM 1sec HGT product with a resolution of 30 meters is selected for the dataset. The DEM is provided in the dataset after normalization based on the maximum elevation of the studied area. The SAR images acquired by the Sentinel-1 constellation are the VV and VH polarizations of the GRD products acquired in IW swath mode during descending passes. They are orthorectified with respect to the DEM to be projected in the ground geometry with a resolution of 10 meters. Similar to the two previous datasets, the long-tail distribution of SAR values is compressed by applying a logarithmic transformation. By their side, the optical images which compose the dataset are the Level-2A products obtained from the Sentinel-2 constellation with a cloud cover below 25%. For the same reasons as for BigEarthNet-MM, the optical and SAR images are selected as temporally close as possible. The RGB channels are selected and masked with the provided cloud and snow maps to ensure reliable ground correspondences. All masked pixels are set to 0 to be easily identifiable. Finally, the intensities are linearly transformed channel by channel from 0 to 255. The images are then cropped to generate 53,475 triplets of patches with a size of 256×256px and with a resolution of 10 meters in ground geometry.

This dataset consists of images acquired under the typical configuration encountered in slope instability detection. Indeed, it is located in a rural and seasonally variable region with slopes steep enough to cause strong distortions in radar images. Figure 4.1 illustrates the strong foreshortening and shadow effects on a radar image extracted from the dataset.

[4]https://ieee-dataport.org/documents/sar-dem-optical-mountainous-dataset-distortion-management, last visited on June 28, 2024

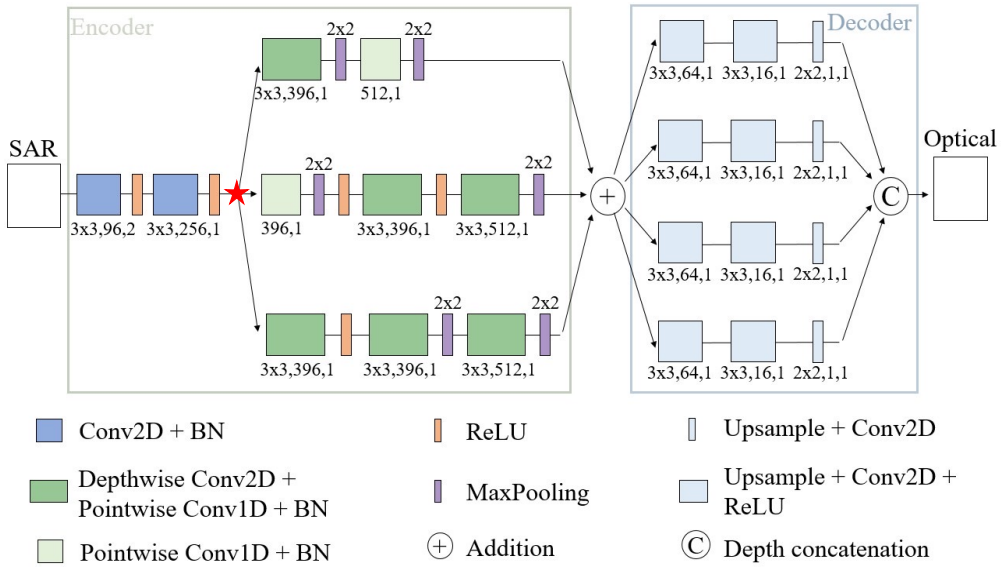


Figure 4.2: SARDINet architecture. Each convolutional block is associated with the size s and number N of the kernels as well as the stride st as "sxs, N, st". The size of MaxPooling windows is also denoted on top of each layer. The red star distinguishes the start of the multi-scale feature extractor.

4.3 Distortion Robust Modality Translator

This section introduces our main contribution in terms of modality translation. The SAR Distorted Image translator Network (SARDINet) proposed in [Bralet et al., 2022] is designed to address distortion issues in SAR images and to reconstruct optical images without unreliable artifacts. In the following, the architecture is introduced and motivated before presenting the first results on the subset of SpaceNet 6 [Shermeyer et al., 2020] described in Section 4.2.1.

4.3.1 SARDINet Architecture

The proposed architecture shown in Figure 4.2 is designed under an encoder-decoder configuration. Indeed, after an input conditioning step, the features are passed to a multi-scale feature extractor. The final optical translation is reconstructed channel-wise by inputting the extracted features into each channel-specific branch of the decoder. More details on the three components are given below.

Input Conditioning: The very first layers of the encoder are aimed at maintaining the most relevant information from the image. They are used not only to reduce the resolution of the image, but also to filter the speckle noise and keep only relevant patterns as input to the multi-scale feature extractor. Since SARDINet is a fully convolutionnal network, it is independent from the size of the input image. Only the number of input channels of the input layer should be updated depending on the available SAR channels. The extraction is done by a succession of two 2D convolutional (Conv2D) layers. The first applies a learned downsampling function by using a stride of 2, and the second focuses on extracting relevant features from the remaining noise with 256 output channels. Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation are introduced after each convolutional step to facilitate the convergence of the training.

Multi-scale feature extractor: This component aims to extract different features from the same input through three distinct branches. To ensure the extraction of spatial and channel-specific features, each branch is composed of Separable Convolution (SepConv) blocks as proposed in the Xception network in [Chollet, 2017]. These blocks - represented by the dark green boxes in Figure 4.2 - successively apply depthwise and pointwise convolutions to the input features. Depthwise convolutions separate the channels of the input features and pass each of them through an independent Conv2D layer to ensure the extraction of spatial-

specific features. The latter are then concatenated to be fused together by the pointwise convolutional layer. This layer consists of applying a 1D convolutional layer (Conv1D) along the channel dimension which can be associated to a learned linear combination of the different channels. This step targets the computation of channel-specific features based on the spatial ones. As visible in Figure 4.2, the three branches of the feature extractor differ in the depth of the extraction applied to each resolution. The first branch (on top of Figure 4.2) aims at extracting contextual features. Indeed, it standardly applies a succession of SepConv blocks, BN and MaxPooling layers to reduce the resolution of the features. The last two branches are dedicated to the extraction of specific features by applying more SepConv blocks and thus ensuring a deeper and more precise feature extraction depending on the resolution. Thus, the second and third branches extract more specific features at low and medium resolution, respectively. These branches capture the structural properties of the SAR images and provide them to the decoder for structurally consistent reconstructions. The latent features result from the sum of the features extracted by the three branches.

Channel independent decoder: Based on the number of optical channels to be reconstructed, the decoder is composed of the same number of independent branches, each one being dedicated to the reconstruction of a single channel. Figure 4.2 illustrates the case where four channels are reconstructed - typically RGB-NIR images from SpaceNet 6. The main reason for such a channel separation is its flexibility in terms of target output and its easy transferability. Indeed, this architecture firstly allows SARDINet to be trained on multiple datasets with different numbers of optical channels. Second, for future fine-tuning purposes, one can leverage the training of SARDINet on a large dataset (*e.g.* BigEarthNet-MM [Sumbul et al., 2021] or SEN12MS [Schmitt et al., 2019]) to ensure high performance in RGB channels and add a new decoder branch to reconstruct an additional channel without having to retrain the three RGB branches from scratch. Each branch consists of applying three blocks of bilinear upsampling, Conv2D layer and ReLU activation. Bilinear upsampling is preferred over transposed convolutions because it has shown better performance than transposed 2D convolutional layers - more insight into these issues in Section 4.4.2.

The architecture of SARDINet is thus designed to limit the impact of the speckle noise on SAR images, as well as to deal with geometrical distortions. Indeed, the structural features extracted at multiple resolutions prevent the network from focusing only on the bright and salient layover pixels and force it to explore more variable characteristics to learn more accurate correspondences between structural SAR and radiometrical optical patterns. In addition, the first branch which targets global feature extraction, helps to take the overall context into account and to reconstruct the most likely and coherent optical patterns in shadow areas. Finally, SARDINet is also designed to be easily adaptable to the reconstruction of additional channels. Indeed, the channel-wise decoder allows to fine-tune already trained channels but also to add more branches and reconstruct new channels without having to retrain the whole network.

4.3.2 Results

4.3.2.1 Experimental setup

For better readability, the experimental setup and the hyperparameter configurations are summarized in a table for each of the experiments performed in this thesis. In particular, the configuration used to train SARDINet on SpaceNet 6 is detailed in Table 4.1. The specificity of the configuration lies in the data augmentation strategy. Indeed, to limit potential overfitting, five random crops are extracted from each of the input images and fed to the network. Since the cropping step is reinitialized for each epoch, the training dataset is thus composed of 1,330 variable samples extracted from the same 266 original images. For comparison, the two standard approaches from the literature - namely Pix2pix [Isola et al., 2017] and CycleGAN [Zhu et al., 2017] - are trained with the same configuration as SARDINet except for the losses which remain identical to the original works. Their training is performed with the PyTorch implementation available on Github[5]. For the sake of overfitting verification, the training of SARDINet is also extended to 400 epochs. The training of SARDINet is performed with a NVIDIA RTX A3000 GPU which computes 100 epochs within an hour.

[5]https://github.com/junyanz/PyTorch-CycleGAN-and-pix2pix, last visited on June 28, 2024.

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	SpaceNet 6	Loss:	MSE	Metrics:	RMSE, MAE, SSIM
Input:	SAR (HH, HV, VH, VV)	Optimizer:	Adam		
Output:	Optical (RGB-NIR)	Batch Size:	32	SOTA:	Pix2pix [Isola et al., 2017]
Size:	200×200 px	Learning Rate:	1.10^{-3}		CycleGAN [Zhu et al., 2017]
#Training:	1330*	#Epochs:	200		
#Evaluation:	165				
#Test:	165				
Specificities					
* Training images are randomly cropped at each epoch - thus the training dataset is variable.					

Table 4.1: SARDINet training configuration. SOTA is standing for State-Of-The-Art.

4.3.2.2 Results and discussion

Quantitative and qualitative results on SpaceNet 6 patches are presented in Table 4.2 and Figure 4.3. According to the former, the performance of SARDINet significantly exceeds that of the literature in terms of quantitative metrics. Several factors can justify this behavior. First of all, the reliability of the reconstructions is greatly improved by SARDINet. This is certainly due to the fact that it is not adversarially trained, which prevents from the generation of hallucinated artifact. The CycleGAN reconstruction in the first row of Figure 4.3 illustrates such hallucinated artifacts since it generates two fake tanks at the bottom of the image. Second, the proposed architecture succeeds in accurately reconstructing the smallest patterns such as the tiny tanks in the second row of Figure 4.3. Such a behavior is made possible thanks to the multi-scale branches which are assigned a specific resolution, thus preventing the loss of tiny but important patterns. This conclusion is even more supported by the results obtained by SARDINet after 400 training epochs where the tanks are better delineated in the fifth column of Figure 4.3. Thirdly, instead of focusing on image credibility, SARDINet ensures structural consistency. Indeed, the SSIM score highlights that SARDINet better translates the structural properties of the SAR image into the optical modality despite an apparent lack of contrast. In particular, the circular shape of the tanks, the shadows but also the straight streets are better reconstructed and less distorted than in the reconstructions of Pix2pix and CycleGAN. This behavior highlights a much better robustness to the input SAR geometrical distortions than adversarial approaches. Finally, the training of SARDINet is still tunable. The increase in performance of the network trained for 400 epochs shows not only quantitatively but also qualitatively that the apparent drawbacks of the methods can be mitigated with a longer and fine-tuned training. In particular, the blurriness visible in SARDINet results is greatly reduced and the areas with lower intensities are appearing in the reconstructions of the fifth column of Figure 4.3.

Network	↓ RMSE	↓ MAE	↑ SSIM-3	↑ SSIM-11
CycleGAN [Zhu et al., 2017]	0.163	0.102	0.780	0.800
Pix2pix [Isola et al., 2017]	0.155	0.098	0.79	**0.814**
SARDINet	**0.080**	**0.053**	**0.850**	0.813
SARDINet 400	0.074	0.046	0.860	0.825

Table 4.2: Quantitative comparison of our methods with the state of the art. Best 200 epochs results are displayed in bold - SARDINet 400 is kept aside from the comparisons.

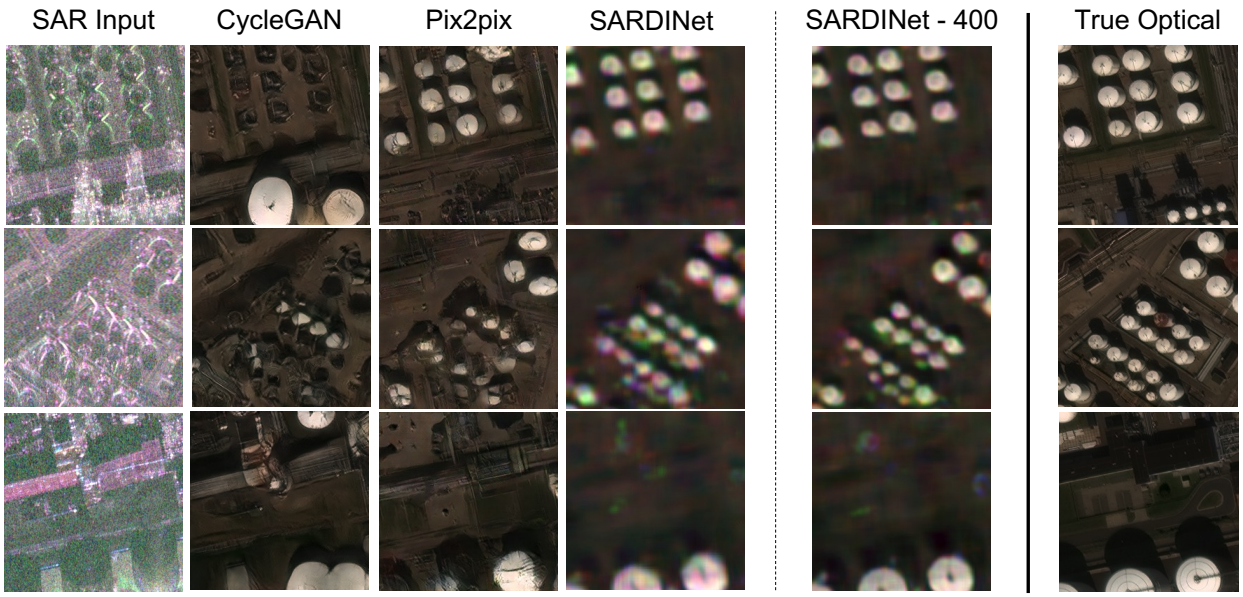


Figure 4.3: Qualitative performance of SARDINet compared to the adversarial approaches from the literature.

4.3.3 SARDINet Outcomes

According to the previous results and discussions, the implementation of SARDINet demonstrates strong capabilities in translating a SAR image into its optical counterpart. In particular, the SAR distortions from the input are successfully overcome by SARDINet while standard adversarial networks from the literature suffer from distorted optical reconstructions. Thus, despite the apparent lack of contrast in SARDINet reconstructions, it achieves the best performance and the most accurate qualitative results. Yet, since SARDINet is intended to be used for slope instability collapse detection, further studies are conducted in the following to confirm its applicability in rural and especially in mountainous regions.

4.4 Architecture Improvements

The studies and experiments proposed in this section aim at extending SARDINet in a rural and/or mountainous context. Due to the large variety of patterns in such regions, this task requires more attention as transfer learning from urban to rural regions would suffer from domain shift issues. Therefore, in this section, two architectural modifications are proposed to adapt SARDINet to a rural context and to strengthen the relevance of the reconstructions. First, the modifications are applied to the input conditioning of SARDINet to integrate additional information as input. Then, different decoding strategies are explored to evaluate their impact on the translations.

4.4.1 Fusion Strategy

This section focuses on improving the performance of SARDINet by studying the impact of its input conditioning fusion strategy when including a second input. Indeed, in mountainous areas, the Digital Elevation Model (DEM) can be fed to the network to strengthen its understanding of the underlying area from a topographic point of view. In particular, the DEM is directly correlated to the imaged slopes which are the source of geometrical distortions in SAR images. Thus, inputting the DEM into the network aims to improve the filtering and the extraction capabilities of its input conditioning for more reliable translations. This section details the works that explore different configurations to incorporate the DEM as a second input of SARDINet and which are proposed in [Bralet et al., 2024c].

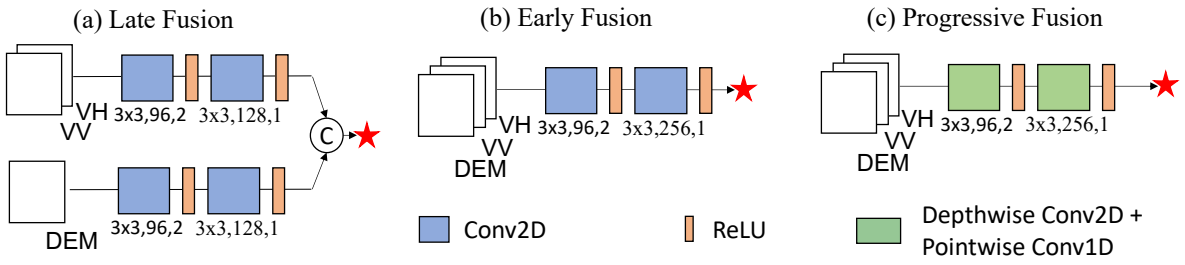


Figure 4.4: Fusion strategies explored for integrating the DEM information in the input conditioning of SARDINet. The red star refers to the input of the multi-scale feature extractor in Figure 4.2.

4.4.1.1 Fusion strategies

This work proposes to investigate three different DEM-SAR fusion strategies. The two standard late and early fusions are explored to assess whether the inputs should be processed independently or jointly, respectively. A third lightweight approach is proposed by using separable convolutions to fuse the inputs progressively throughout the input conditioning while reducing the size of the network. The three strategies are as follows.

Late Fusion: The strategy in Figure 4.4a considers the SAR image and the DEM as two separate inputs. Thus, the input conditioning is split into two different branches to process them independently. The features extracted from both inputs are then concatenated and fed to the multi-scale feature extractor of SARDINet. Therefore, late fusion focuses on extracting sufficiently informative features in both branches and leaves the fusion step to the multi-scale feature extractor. Since the latter consists in SepConv blocks (see Figure 4.2), the DEM extracted features can be progressively integrated with the SAR features.

Early Fusion: By concatenating the DEM with the SAR image, early fusion proposes a three-channel image to the input conditioning of SARDINet as illustrated in Figure 4.4b. Such a strategy allows the network to integrate the DEM for the SAR filtering and feature extraction processes. Therefore, the input conditioning is not only responsible for despeckling but also for the fusion of the modalities to focus the feature extraction only in relevant areas. The main drawback of this approach is that the raw DEM information is only used in the very first Conv2D layer while the rest of the network focuses on feature refinement.

Progressive Fusion: To facilitate the transfer of DEM features to the deeper layers, the third strategy leverages SepConv blocks [Chollet, 2017] in Figure 4.4c. As in the early fusion strategy, the modalities are concatenated as input to the network. However, instead of using standard Conv2D layers that would fuse SAR and DEM features together, SepConv blocks are applied to delay this fusion step. Indeed, the depthwise convolutions aim at extracting modality-specific features which are then merged together by using the pointwise convolutions. Therefore, the DEM features can be propagated further in the network and be integrated more progressively than in the early fusion strategy. Meanwhile, the input conditioning can still leverage the DEM for the SAR feature extraction, unlike the late fusion strategy. In the original SepConv blocks [Chollet, 2017], a single spatial kernel is applied to each input channel which in the current configuration would result in a single spatial extraction per modality. This is avoided in this strategy by applying a kernel balancing in between the depthwise and pointwise convolutions as depicted in Figure 4.5. First, multiple spatial kernels are extracted from the C_{in} input channels in the depthwise convolutions. The K_S spatial feature maps of each channel are then concatenated along the channel dimension to provide K_S features with C_{in} channels to the pointwise convolution. The latter combines each input feature with the same K_C kernels to extract inter-channel characteristics. Finally all these features are concatenated to output $K_S \times K_C$ feature maps.

Each implemented strategy may theoretically have advantages over the others, so the objective of the study is thus to evaluate which of them best exploits the DEM to extract actually relevant SAR features leading to more accurate translations. For the sake of conciseness and readability of the following discussion, the Late, Early and Progressive strategies are referred to as SARDINet$_L$, SARDINet$_E$ and SARDINet$_P$.

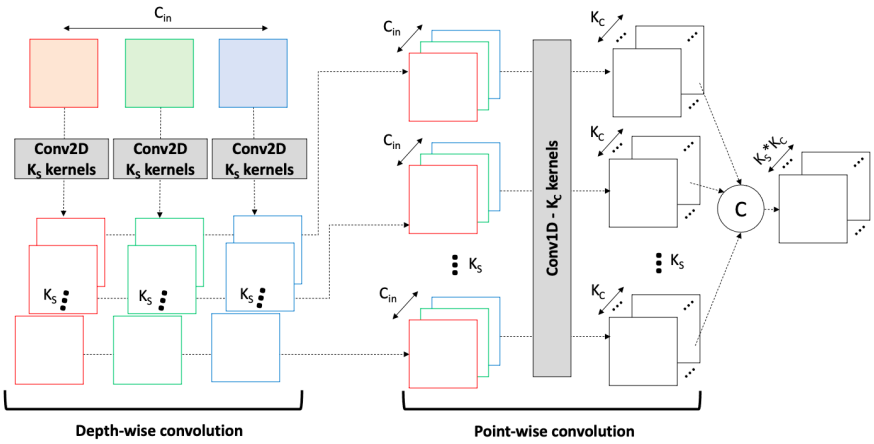


Figure 4.5: Balanced separable convolutions allowing the extraction of more spatial features.

4.4.1.2 Experimental setup

The training of these different fusion strategies on the mountain dataset of Section 4.2.3 is performed according to the configuration in Table 4.3. In particular, the calculation of the MSE, SSIM and PSNR values is done on the whole image but also on specific distorted areas. The masks of the distorted regions are obtained using the approach developed in [Meier et al., 1993] with the Sentinel Application Platform (SNAP)[6]. This distinction between distorted and clean regions allows for a better understanding and evaluation of the ability of each strategy to infer the most likely land-cover despite the geometrical distortions. The FID score is computed only over the whole image since it consists in studying the behavior of an alternative Inception network [Szegedy et al., 2016]. Masking the input to keep only distorted areas would not be informative with this metric as the Inception network would focus on the masked areas rather than on the actual distorted region of interest. To be consistent with Section 4.3, both Pix2pix [Isola et al., 2017] and CycleGAN [Zhu et al., 2017] are trained to reconstruct an optical image from the SAR-DEM input pair from the mountain dataset. In addition, a SAR-only version of SARDINet is trained to better evaluate the impact of the DEM on the reconstructions. These three architectures are trained with the same configuration as in Table 4.3 - except for the losses of the adversarial networks which remain unchanged. To study the stability of the different networks, the performance are averaged over five trainings with a different random initialization. SARDINet variants have been trained with a NVIDIA A6000 GPU which computes 100 epochs in 13 hours. It is worth mentioning that the architectural modifications have a low additional computational cost from 15 minutes to 1 hour.

4.4.1.3 Results

The performance of each network is detailed quantitatively in Table 4.4 and qualitatively in Figure 4.6. The goal of this study is threefold: first, to assess the impact of the DEM on the SAR-to-optical translations, then to find the best way to fuse it with the SAR input and finally, to evaluate whether SARDINet still outperforms the state-of-the-art methods in rural areas.

DEM impact: The comparison of the different fusion strategies with the original SARDINet trained only with the SAR image clearly shows advantages for all metrics except for the FID score. Thus, including the DEM as an input to the network seems to widen the gap between credible and relevant pattern reconstruction. Yet, the goal of this section is to improve the relevance of the reconstructed patterns by including the DEM before improving the overall credibility of the image (see Section 4.5). In addition, the behavior of the DEM-assisted networks on the distorted areas significantly improve the performance of the original SARDINet not only quantitatively but also qualitatively. Indeed, one of the major challenges related to distorted areas

[6]https://step.esa.int/main/toolboxes/snap/, last visited on June 28, 2023

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	Mountain Dataset	Loss:	MAE	Metrics*:	MSE, SSIM, PSNR, FID**
Input:	SAR (VH, VV) + DEM	Optimizer:	Adam		
Output:	Optical (RGB)	Batch Size:	24	SOTA:	Pix2pix [Isola et al., 2017]
Size:	256×256 px	Learning Rate:	1.10^{-5}		CycleGAN [Zhu et al., 2017]
#Training:	42.780	#Epochs:	100		SARDINet† [Bralet et al., 2022]
#Evaluation:	5.348				
#Test:	5.347				
Specificities					
* Metrics are applied on the whole images but also on distorted areas only. Metrics are averaged over 5 training with different configurations.					
** The FID is only applied on the entire image.					
† The training of SARDINet is only performed with the SAR image as input - its DEM counterpart being SARDINet$_E$.					

Table 4.3: Fusion strategies training configuration. SOTA is standing for State-Of-The-Art.

is the shadows which can be confused with water bodies. Since water often acts as a mirror, it does not backscatter a signal to the satellite. Therefore, in the SAR images, the lake visible in the third row of Figure 4.6 is radiometrically similar to the red circled shadow area in the first row. Based on the translations, both SARDINet and its DEM derivatives successfully reconstruct the lake, but the former confuses the shadow regions with a water body and reconstructs a river-like pattern in the red circled region. Conversely, the confusion is strongly attenuated in the results of SARDINet$_E$ and avoided with SARDINet$_L$ showing that the DEM is appropriately leveraged by the networks to obtain more reliable reconstructions.

Comparison of fusion strategies: From a quantitative point of view, SARDINet$_L$ and SARDINet$_E$ show a very similar performance, except for the FID score where the former shows a decrease of 0.16 points. Nevertheless, from a qualitative point of view, more interesting conclusions can be drawn. First, as already mentioned, in the first row of the Figure 4.6, SARDINet$_E$ still shows a river-like pattern in the red circled shadow area. In addition, the yellow circles in the third row show that SARDINet$_E$ reconstructs more unreliable urban-like patterns than its competitors. Both anomalies illustrate that fusing the DEM early in the network may limit the extraction of the most appropriate DEM features and allow the SAR image to dominate the feature extraction process because it contains more variable information. Yet, forcing the network to extract DEM features in a separate branch helps to better understand the contextual features of the area and to prevent one input from dominating the other. The progressive fusion SARDINet$_P$ was also implemented to avoid such behavior. However, as reflected by both the quantitative and qualitative performance, it achieves the worst results of all - even worse than the original SARDINet. The main reason for this drop in performance is the large reduction in the number of parameters: 9,852 parameters against more than 220,000 for the three other SARDINet-like architectures. Despite this weight reduction, SARDINet$_P$ still shows competitive performance, which encourages for further research towards lightweight networks.

State-of-the-art comparison: Finally, the comparison of the SARDINet architectures with the standard methods from the literature illustrates the strong and stable capabilities of the formers. Indeed, although Pix2pix and CycleGAN show the best FID results due to their adversarial training, they suffer from a critical lack of relevance both quantitatively and qualitatively speaking. The results of the other three metrics in Table 4.4 show a strong performance drop compared to the proposed approaches not only in the distorted areas but also on the whole image. In addition, their results are much less stable than SARDINet as depicted by the standard deviation - even for the FID score. Figure 4.6 highlights some of the reasons for these poor results. On the one hand, despite its apparent credibility, the training of Pix2pix seems to be sensitive to the presence of clouds in the dataset which implies the generation of cloud patterns in the reconstructions. More critically, the lake in the image of the last row is not reconstructed at all by the network. On the other hand, CycleGAN manages to reconstruct less distorted patterns. Yet, it is sensitive to the land-cover

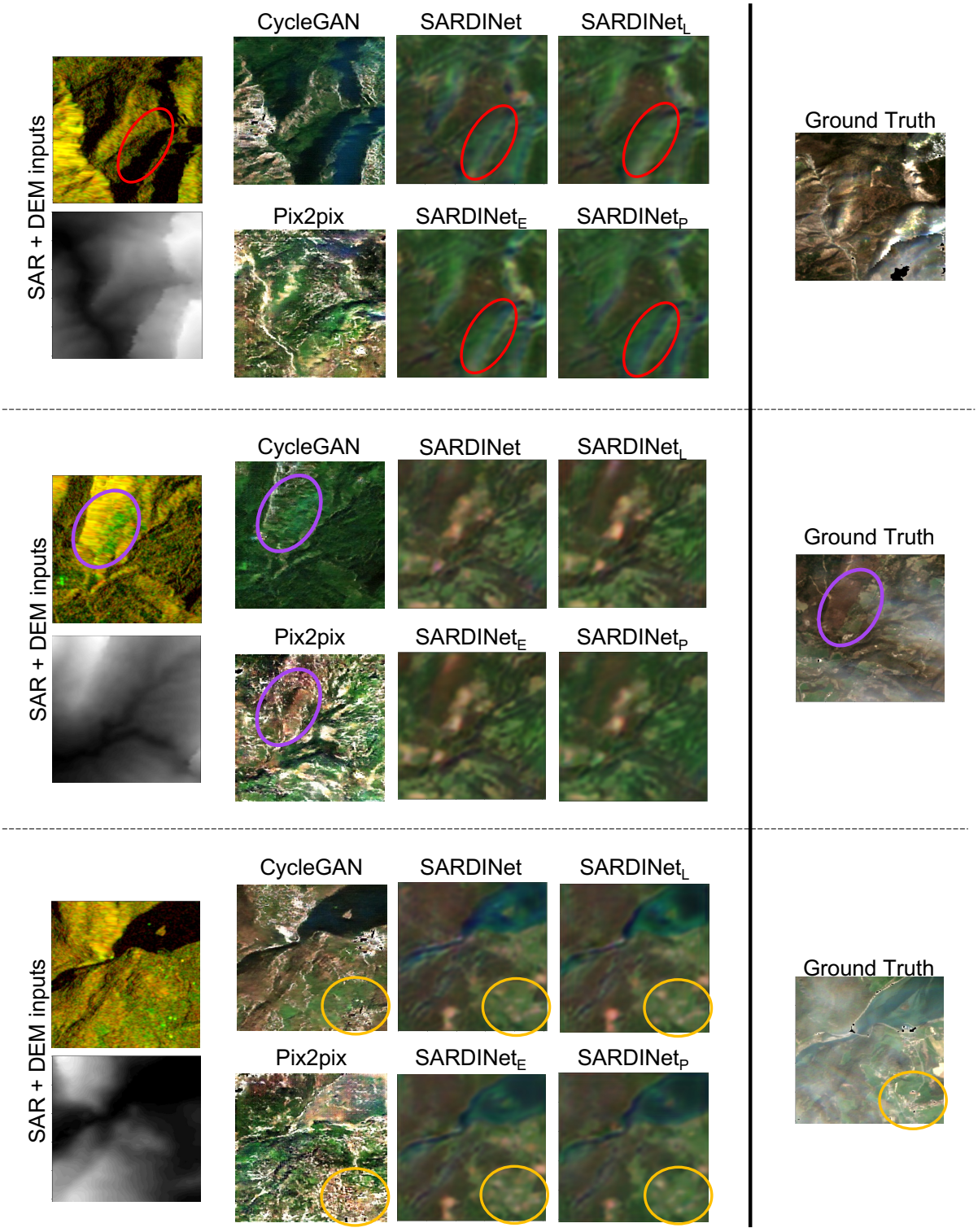


Figure 4.6: Qualitative performance of the fusion strategies applied on SARDINet compared to the state of the art on mountain regions.

	↓ MSE ($.10^{-2}$)	↑ SSIM ($.10^{-2}$)	↑ PSNR	↓ FID	↓ MSE ($.10^{-2}$)	↑ SSIM ($.10^{-2}$)	↑ PSNR
	Global				Foreshortening areas		
CycleGAN	8.67±2.10	61.11±4.49	27.99±0.06	**10.67**±13.14	14.22±2.89	57.95±2.18	26.06±0.03
Pix2pix	8.19±1.27	59.14±1.46	28.03±0.04	21.71±8.67	8.05±1.12	59.94±1.91	26.25±0.07
SARDINet	2.91±0.06	74.44±0.07	28.58±0.04	38.73±0.54	3.41±0.12	74.78±0.11	**26.99**±0.09
SARDINet$_L$	2.82±0.06	**74.48**±0.07	**28.59**±0.05	38.94±0.51	**3.20**±0.08	**74.89**±0.15	26.97±0.07
SARDINet$_E$	**2.81**±0.04	74.47±0.02	**28.59**±0.03	39.10±0.70	**3.20**±0.04	74.84±0.08	26.98±0.10
SARDINet$_P$	2.99±0.04	74.26±0.06	28.55±0.03	34.01±0.54	3.40±0.05	74.71±0.04	26.97±0.07
	Non distorted areas				Shadows areas		
CycleGAN	8.42±2.19	61.27±4.62	27.99±0.07	*NA*	15.33±0.35	55.34±3.12	25.68±0.01
Pix2pix	8.19±1.29	59.12±1.45	28.03±0.04	*NA*	12.27±1.60	55.24±1.18	25.79±0.01
SARDINet	2.84±0.06	74.53±0.07	28.58±0.04	*NA*	6.16±0.16	66.92±0.06	26.04±0.02
SARDINet$_L$	2.76±0.06	74.55±0.07	**28.60**±0.05	*NA*	5.85±0.14	67.00±0.05	**26.05**±0.02
SARDINet$_E$	**2.75**±0.05	**74.56**±0.02	28.59±0.03	*NA*	**5.80**±0.09	**67.02**±0.05	26.04±0.02
SARDINet$_P$	2.92±0.04	74.34±0.06	28.55±0.03	*NA*	6.18±0.14	66.65±0.09	26.01±0.02

Table 4.4: Quantitative comparisons of the fusion methods with the literature depending on SAR distortions. Mean and standard deviation over 5 experiments with different random initializations are displayed. Best results are displayed in bold.

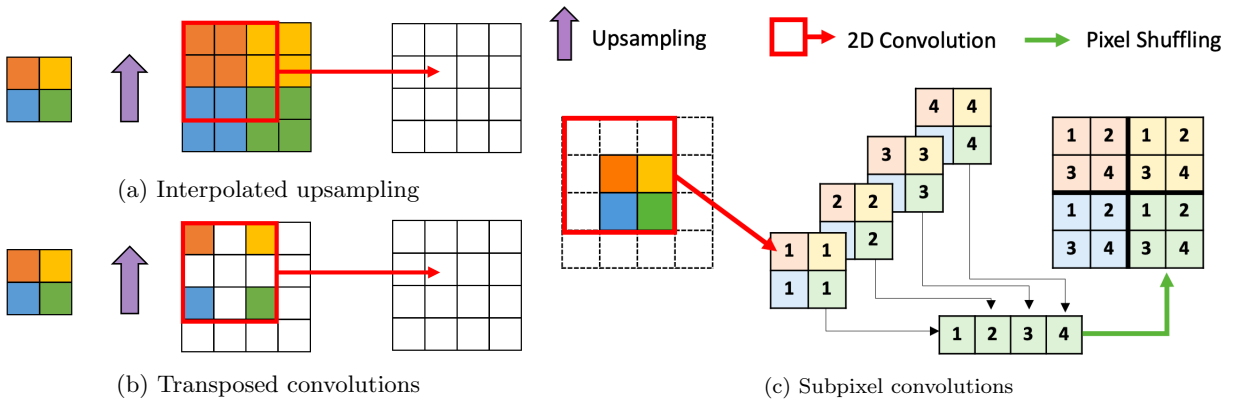
distribution in the training dataset. Indeed, as visible in the first two rows of Figure 4.6, forest-like patterns are reconstructed over the entire image. In addition, the shadow areas of the image in the first row are left unchanged or colored blue - showing the inappropriate integration of the DEM in the feature extraction. Due to their respective drawbacks, the two approaches are not able to distinguish the separation between the forest and the field in the purple circled area nor the village in the yellow circle in the third row.

4.4.1.4 Fusion strategies outcomes

Two main outcomes can be drawn from this study. On the one hand, the performance of SARDINet in mountainous regions is still high compared to the state of the art. Indeed, the relevance of the reconstructed patterns allows for a better understanding of the area and thanks to its generator-only architecture, it avoids the generation of unreliable artifacts except in highly distorted areas - *e.g.* shadow areas can be confused with water bodies. On the other hand, the integration of the DEM as an input to the network can be appropriately leveraged to better address the distortion issues and reconstruct even more reliable optical patterns in the final image. Therefore, this study brings new possibilities to combine more inputs with the SAR images. Indeed, in addition to the DEM, other data sources could be provided to the model. Among other possibilities, the layover and shadow masks could increase the distortion awareness, a SAR-optical pair from the past could be informative for better style transfer capabilities, or a high resolution SAR image (*e.g.* from TerraSAR-X) could be used to enrich the structural understanding of the area. Finally, the current study focuses on the fusion strategy but due to its low resolution (30 meters for SRTM 1sec HGT), the DEM could be considered as a conditionnal input. Thus, Conditional Batch Normalization [Mescheder et al., 2019] could be implemented to use such information even more appropriately.

4.4.2 Decoding Strategy

Although the encoding strategy ensures the extraction of the most relevant features from the SAR image, the decoding strategy is responsible for generating the final output image. In particular, the creation of high resolution details, the coloration and the contrast are the key roles of the decoder which aims to generate the most probable optical image based on the features extracted by the encoder. Thus, a study of the impact of the decoder strategy on the final reconstruction is conducted in this thesis [Bralet et al., 2023].



(a) Interpolated upsampling

(b) Transposed convolutions

(c) Subpixel convolutions

Figure 4.7: Decoding strategies explored to enhance the relevance of the reconstructed patterns in the decoder of SARDINet.

4.4.2.1 Decoding strategies

Among the wide variety of encoder-decoder architectures implemented in the literature, for instance for modality translation, the decoder is rarely the core of the study. In particular, they are usually designed as a succession of upsampling and convolutional layers to generate high-resolution images from low-resolution features. Nevertheless, the choice of one or another method to reconstruct the output image can strongly modify its quality both quantitatively and qualitatively. This study focuses on exploring the best way to upsample the features from one resolution to another. The most common upsampling stages consist of interpolation layers or in transposed convolutional layers. Another approach introduced in [Shi et al., 2016] is the sub-pixel convolutional layer and consists in shuffling the pixels of the low-resolution features to generate the high-resolution ones. More details on each of the strategies illustrated in Figure 4.7 are given below.

Interpolated upsampling: This strategy simply applies an upsampling operation based on a chosen interpolation method. Nearest neighbor, bilinear and bicubic interpolations are of standard use in the literature for upsampling low-resolution features. Then, to correct susceptible errors due to the upsampling steps (*e.g.* the pixelation of the nearest neighbor approach or the blurriness of the bilinear one), a 2D convolutional layer is often successively applied to the upsampling step. Figure 4.7a illustrates a typical use of interpolated upsampling with the nearest neighbors method.

Transposed convolutions: This strategy relies on 2D convolutions to find the most relevant interpolation to apply to each high-resolution pixel. Indeed, the low-resolution features are first dispatched on the final high-resolution grid - undefined values are set to 0. Then a 2D convolutional layer is applied to the upsampled features to fill in the undefined pixels. This is the preferred strategy nowadays for reasons of experiment reproducibility. Figure 4.7b illustrates the transposed convolution approach.

Sub-pixel convolutions: Contrary to the other two strategies, it consists of applying a 2D convolutional layer directly to the low-resolution features. The goal of the convolution is to apply enough kernels to fill the upsampled neighborhood of each pixel after shuffling. Figure 4.7c describes the strategy within a scheme while an algorithmic description follows. Suppose the upsampling stage aims to multiply the resolution of the features by k. The sub-pixel convolution strategy first applies a 2D convolutional layer with k^2 kernels to the low resolution features. Then the k^2 feature maps are rearranged to reconstruct the neighborhood of each original pixel. To this end, the value of the pixel $p^m_{i,j}$ at the location i, j of the mth feature map is associated with the pixel $p_{ki+q,kj+r}$ of the high-resolution feature - by considering q and r to be the quotient and the remainder of the Euclidean division of m by k. The network is thus free to determine the interpolation that best fits each local feature. The main drawback of this approach is the risk of generating checkerboard artifacts. To avoid this behavior, the authors of [Aitken et al., 2017] suggest to identically initialize the weights of the 2D convolutional layer used to compute the interpolation.

In order to better distinguish the different strategies applied to the decoder of SARDINet, the follow-

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	BigEarthNet-MM	Loss:	MAE + SSIM	Metrics:	MSE, PSNR, FID
Input:	SAR (VH, VV)	Optimizer:	Adam		
Output:	Optical (RGB)	Batch Size:	8	SOTA:	SARDINet [Bralet et al., 2022]
Size:	120×120 px	Learning Rate:	5.10^{-5}		
#Training:	156.087	#Epochs:	100		
#Evaluation:	39.022				

Table 4.5: Decoding strategies training configuration. SOTA is standing for State-Of-The-Art.

ing notations are proposed: SARDINet$_{up}$ for the interpolated upsampling strategy, SARDINet$_{tr}$ for the transposed convolution strategy and SARDINet$^{pix}_{pix}$ for the sub-pixel convolution strategy. In addition, two more variants are proposed by combining the three previous ones. Indeed, in the vast majority of works, the upsampling stages are kept identical throughout the decoder. Nevertheless, sub-pixel convolutions were originally designed for the super-resolution task, which is different from the feature reconstruction task of the first decoding layers. Therefore, SARDINet$^{pix}_{up}$ and SARDINet$^{pix}_{tr}$ are implemented by applying a sub-pixel convolutional layer only for the very last upsampling stage of the decoder, the previous ones being composed of interpolated upsampling and transposed convolutional layers, respectively.

4.4.2.2 Experimental setup

The overall configuration of the experiments can be found in Table 4.5. It is noticeable that for these experiments, only SARDINet variants are trained on BigEarthNet-MM without any comparison to the state-of-the-art. Indeed, previous studies have already compared the performance of SARDINet with the literature in both urban and rural contexts including SAR distortions. Furthermore, the study conducted in this section does not aim to enhance the contrast but only the relevance of the reconstructed patterns through the decoding strategy. Therefore, since the goal is to improve SARDINet, the latter is the only reference of this study. Based on the original SARDINet architecture composed of bilinear upsampling layers, the same interpolation is chosen for this work and therefore, SARDINet$_{up}$ refers to SARDINet hereinafter.

From an implementation point of view, the number of kernels of each decoding convolutional layer and the number of upsampling stages composing the different SARDINet variants are identical to ensure fair comparisons. The only difference is that all the networks with a final sup-pixel convolutional layer (denoted as SARDINet$^{pix}_{xx}$ where xx refers to the hidden layer upsampling strategies) output the final image directly while the others require an additional 2D convolutional layer. Finally, the kernels of the sub-pixel convolutions are randomly initialized to avoid any bias in the training. It is worth mentioning that none of the implementations impacted the training computational time.

4.4.2.3 Results

The quantitative and qualitative results of the different decoding strategies of SARDINet are presented in Table 4.6 and Figure 4.8, respectively. Among the several conclusions that can be drawn from these results, three of them are of particular interest.

The interpolation should be learned: Based on the qualitative results of SARDINet$_{up}$ and SARDINet$^{pix}_{up}$, the reconstructions appear to be more blurred and less reliable than the other strategies. This is also illustrated by the FID score which strongly increases due to the lack of contrast in the images. In particular, in the first image, the boundaries between the forest and field patterns is not distinguishable with SARDINet$_{up}$. Further experiments were performed with different interpolation methods but the same behavior occurred in the final reconstructions. Due to the complexity of the patterns to be reconstructed, the application of standard interpolation upsampling strategies seems to be more harmful than beneficial for the network.

	↓ MSE ($.10^{-3}$)	↑ PSNR	↓ FID
SARDINet$_{up}$	10.69	29.19	<u>6.06</u>
SARDINet$_{tr}$	11.99	29.24	6.21
SARDINet$^{pix}_{pix}$	10.78	**29.34**	**5.39**
SARDINet$^{pix}_{up}$	**9.87**	<u>29.28</u>	6.10
SARDINet$^{pix}_{tr}$	<u>10.55</u>	29.27	**5.39**

Table 4.6: Quantitative performance of the different decoding strategies. Best and second best results are bolded and underlined, respectively.

Therefore, letting it optimizing the interpolation strategy depending on the reconstructed features and on the neighborhood of each pixel with convolutional layers is the most accurate strategy.

The ultimate sub-pixel convolution layer is beneficial: According to the increase in performance of SARDINet$^{pix}_{up}$ and SARDINet$^{pix}_{tr}$ compared to that of SARDINet$_{up}$ and SARDINet$_{tr}$, the use of sub-pixel convolutions for the last layer is highly beneficial for the reconstruction. Indeed, this last layer is specifically aimed at including more detailed patterns in the image without affecting their overall relevance which has already been determined in the previous decoding layers. Thus, this strategy greatly improves the qualitative results: for instance, the transitions between the different land-cover classes are better delineated in the first two rows of Figure 4.8 and some forest patterns are accurately reconstructed in the last row with SARDINet$^{pix}_{tr}$ compared to SARDINet$_{tr}$. All these improvements explain the gain of 0.82 points in terms of FID score but also of 1.44 points in terms of MSE score. Even more interestingly, not only do the sub-pixel convolutions have powerful end-layer capabilities but they can also accurately reconstruct intermediate decoding features. Indeed, based on Table 4.6, SARDINet$^{pix}_{pix}$ achieves the best balance in terms of metrics. Nevertheless, the sensitivity of these layers to the checkerboard artifacts [Aitken et al., 2017] coupled with the cascade of layers used in the decoder strongly affects the final optical image. Indeed, when comparing the results of SARDINet$^{pix}_{pix}$ with SARDINet$^{pix}_{tr}$ in the red circled regions in Figure 4.8, the reconstructed artifacts are more distorting the former ones.

The transposed convolution is dedicated to feature reconstruction only: As for the very first implementation of SARDINet [Bralet et al., 2022], the transposed convolution-only decoder SARDINet$_{tr}$ shows the worst quantitative performance of all in Table 4.6. Nevertheless, the modification of the very last layer into a sub-pixel convolution layer ends up with the second best balance of metrics and - if the checkerboard artifacts are taken into account - the best quantitative-qualitative balance. This behavior demonstrates that transposed convolutions are particularly powerful for the feature reconstruction stage (*i.e.* in the first decoding layers) but that their capabilities in super-resolution are poorer in the last layers. Indeed, the fact to dispatch the low-resolution features to a high-resolution grid (see Figure 4.7b) is especially helpful for the creation of novel features with poor contextual information. However, for the completion and detailing of an already predefined context - *i.e.* in the last layer of the decoder - this dispatching step disturbs the network for the final reconstruction. For this reason, the most accurate and less distorted optical reconstructions are obtained by coupling the creation skills of the transposed convolutions with the super-resolution capabilities of the sub-pixel convolutions.

4.4.2.4 Decoding strategies outcomes

The decoding strategies experimented above demonstrate the importance of the interpolation for the final reconstructions. In particular, the poor performance obtained with the interpolated upsampling layers highlights the fact that the interpolation strategy to be used for upsampling is highly dependent on the features extracted in the previous layers. Therefore, allowing the network to learn the most appropriate interpolation for each case with the transposed convolutions in the low resolution layers helps to reconstruct and generate more relevant features of interest. Finally, the application of sub-pixel convolutions in the last layers shows strong capabilities for the super-resolution of the final image, *i.e.* enhancing the contrast

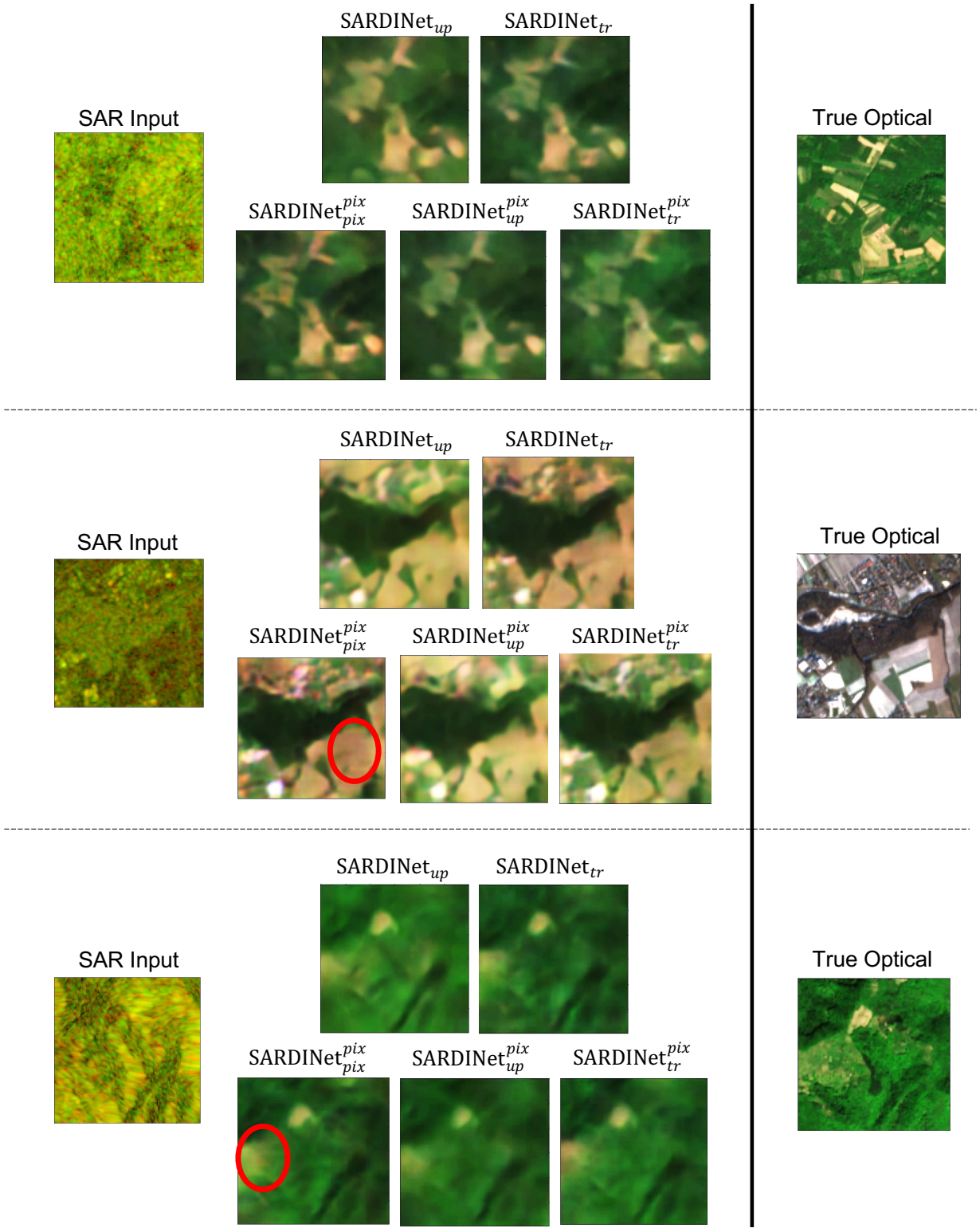


Figure 4.8: Qualitative performance of SARDINet with different decoding strategies.

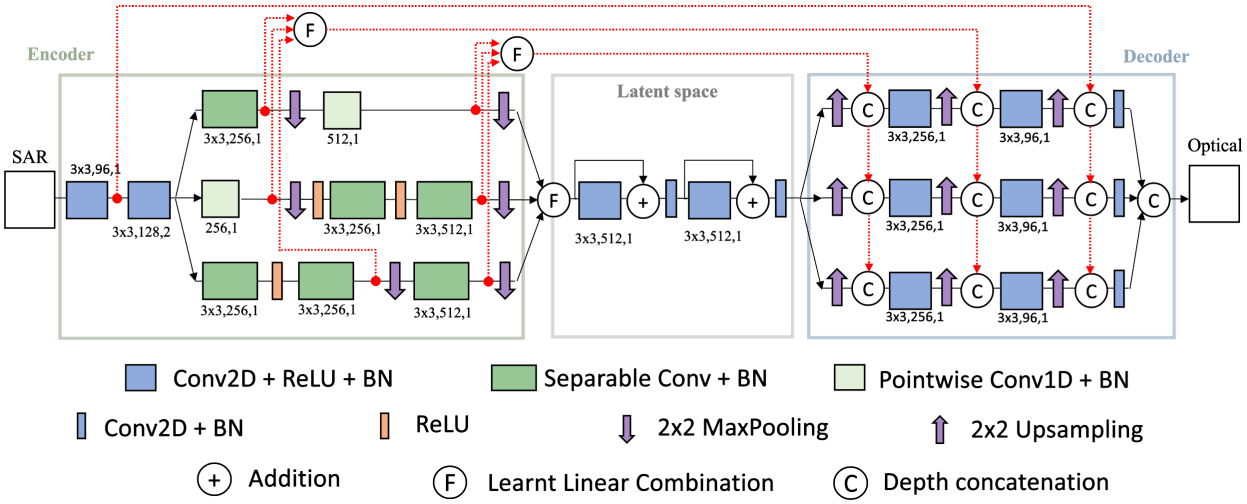


Figure 4.9: Architectural modifications of SARDINet leading to the final SARDINetV2 architecture.

without generating artifacts. Nevertheless, the range of high resolution details reconstructed by the network is still extremely poor which limits the capabilities of the sub-pixel convolutions for contrast enhancement.

4.5 Contrast Enhancement

Following the architectural modifications which aimed at improving the relevance of the reconstructed features, this section describes the strategies which target the contrast enhancement of the final optical image. Thus, two challenges are addressed in this section: improving the credibility of the reconstructed images and maintaining the accuracy of the reconstructed patterns. Indeed, the translations are intended to be used for downstream tasks - in particular, in this thesis, for the detection of sudden slope instability collapses - which implies that the accuracy cannot be traded off against the credibility. To this end, three different strategies are separately investigated to improve the performance of SARDINet:

- Architectural modifications to transfer high-resolution features from the encoder to the decoder,
- The adversarial training of SARDINet to increase the similarity between the true optical and the translated images,
- The application of standard post-deblurring strategies to an already translated image.

The final aggregation of these three intermediate approaches results in a Translation-Deblurring framework that shows promising preliminary results.

4.5.1 SARDINetV2

As experimented in Section 4.4.2, the decoding strategy has an impact not only on the relevance of the reconstructed patterns but also on the contrast and details of the translated image. However, based on the current SARDINet architecture in Figure 4.2, the features transferred from the encoder to the decoder are the low-resolution latent features resulting from the summation of the output of the three branches. Therefore, the decoder should create the most susceptible high-resolution details based on very low-resolution features, *i.e.* should generate patterns and textures without any clue for their actual presence. This constraint significantly makes the task of the decoder more difficult. In this section, the additional architectural modifications proposed in [Bralet et al., 2024b] to reduce the hallucination risk of the decoder are introduced.

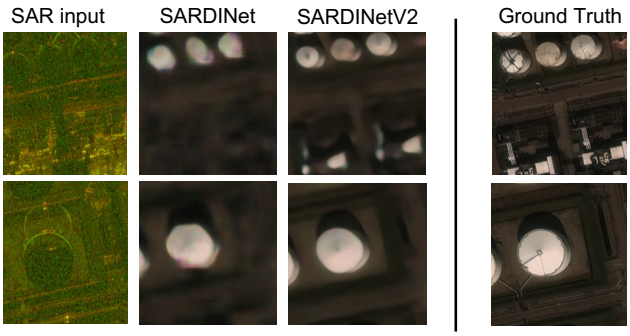


Figure 4.10: Qualitative performance of SARDINetV2 on SpaceNet 6 [Shermeyer et al., 2020].

	SpaceNet 6		
Network	↓ MAE	↑ SSIM-3	↑ SSIM-11
SARDINet	$8.64.10^{-2}$	0.873	0.842
SARDINetV2	$\mathbf{7.25.10^{-2}}$	**0.890**	**0.867**

Table 4.7: Quantitative performance of SARDINetV2 on SpaceNet 6 [Shermeyer et al., 2020]

4.5.1.1 Architecture

Four main modifications are applied to the original SARDINet architecture to improve the reconstruction of details but also the credibility of the reconstructed patterns in the final image. These modifications result in SARDINetV2, which architecture is depicted in Figure 4.9.

1. Since SARDINet was initially designed for urban areas, the geometrical context required less variety in terms of patterns than in rural and especially mountainous regions. Therefore, to allow the network to reconstruct more variable patterns, the first modification consists in balancing the number of kernels between the encoder and the decoder. The convolutional stages in the decoder are made symmetrical to those in the encoder in terms of kernels. This modification trades off the quality and variety of the reconstructed patterns against the memory cost of the network. This trade-off remains acceptable because the use of separable convolutions in the encoder significantly lightens the architecture compared to standard convolutional encoders.

2. For a better SAR-to-optical feature projection, further processing is introduced in the latent space of SARDINet. As illustrated by the gray box in Figure 4.9, two stages of residual blocks [He et al., 2016a] are used for the projection. The residual branches ensure that the structural properties of the SAR image are still transferred to the optical modality.

3. Inspired by the high performance achieved by the U-Net architecture [Ronneberger et al., 2015] and its use in several application areas, the main change introduced in the architecture consists in skip connections. These red dotted arrows in Figure 4.9 allow the transfer of the features extracted at different resolutions from the encoder to the decoder. Thus, this simple modification coupled with the residual blocks in the latent space ensures that the structural properties of the SAR image are appropriately translated into optical patterns. In particular, high-resolution SAR details that are lost during the three downsampling stages of the architecture are still available in the decoder to improve details and contrast in the final optical image.

4. Finally, a learned linear combination is used to appropriately merge the three branches of the encoder before transferring the features through the skip connections. Indeed, in the original architecture in Figure 4.2, the three branches are simply fused in the latent space with the summation function. Nevertheless, due to the multi-resolution aspect of the branches, they do not have the same importance from one stage to another. Using this branch-wise linear combination only adds three weights per branch to the size of the model but ensures that the branches are exploited in the best possible way during feature transfer.

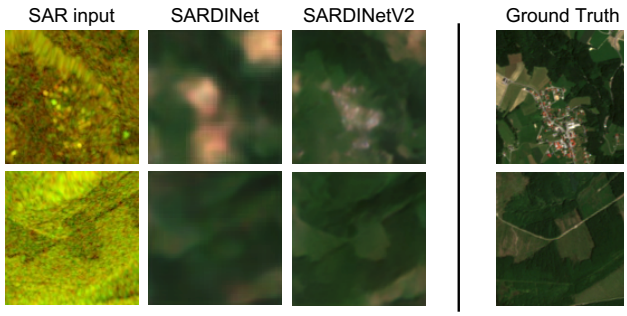


Figure 4.11: Qualitative performance of SARDINetV2 on BigEarthNet-MM [Sumbul et al., 2021].

	BigEarthNet-MM		
Network	↓ MSE	↑ PSNR	↓ FID
SARDINet	$12.51.10^{-3}$	24.82	**4.64**
SARDINetV2	$\mathbf{10.49.10^{-3}}$	**25.26**	5.07

Table 4.8: Quantitative performance of SARDINetV2 on BigEarthNet-MM [Sumbul et al., 2021].

4.5.1.2 Results

Due to the different architectural changes, the performance of SARDINetV2 should be evaluated in two ways: by comparing it to the original version of SARDINet and by verifying that the transfer of high resolution features maintains its distortion robustness. Therefore, SARDINetV2 is compared to SARDINet on two different datasets: SpaceNet 6 [Shermeyer et al., 2020] and a subset of 41,932 samples from BigEarthNet-MM [Sumbul et al., 2021]. The training configuration of the networks on SpaceNet 6 is kept unchanged to Section 4.3.2. On the other hand, the training on the 34,412 samples of BigEarthNet-MM is performed with an Adam optimizer that aims to optimize a linear combination of the MAE and SSIM losses with a learning rate of 5.10^{-5} and a batch size of 8 for 150 epochs. Since this is a training on rural regions, the metrics used to evaluate the remaining 7,520 images are the MSE, PSNR and FID values.

For the sake of reproducibility of the results and due to PyTorch implementations, the upsampling layers of both architectures are based on a nearest neighbor interpolation instead of the bilinear interpolation of the original version of SARDINet. Under such configuration and with a NVIDIA A6000 GPU, SARDINetV2 requires 2 more hours to perform 100 epochs on BigEarthNet-MM compared to SARDINet.

Figures 4.10 and Tables 4.7 summarize the qualitative and quantitative performance of SARDINetV2 compared to SARDINet on SpaceNet 6. The main conclusion to be drawn from these results is that thanks to the architectural modifications, SARDINetV2 significantly outperforms the original SARDINet. Indeed, based on the results obtained on SpaceNet 6, all the metrics are improved with SARDINetV2 by a large margin: more than 1.4×10^{-2} points in MAE and more than 1.5×10^{-2} for both SSIM metrics. Moreover, the qualitative reconstructions are more contrasted than the results of SARDINet. With the attenuation of the blurring aspect obtained by SARDINetV2, the tanks and the buildings are better delineated and the ground patterns are more reliable. In addition, the second row of Figure 4.10 illustrates a case where the tank suffers from strong layovers in the SAR image and where SARDINetV2 reconstructions are still coherent with the ground truth image - showing that the robustness to distortions is maintained despite the architectural modifications.

Then, similar conclusions can be drawn based on the performance obtained in rural regions when trained on the BigEarthNet-MM dataset in Figure 4.11 and Table 4.8. Both relevance metrics show significant improvements when using SARDINetV2 as the MSE score decreases by more than 2×10^{-3} points and the PSNR increases by 0.36 points. Surprisingly, a loss of 0.4 points in FID is observed with the new architecture. Despite such a counter-intuitive performance, the visual results of SARDINetV2 compared to SARDINet in Figure 4.11 still look better contrasted. Indeed, SARDINet is particularly underperforming in the reconstruction of the village in the first row while SARDINetV2 succeeds to find out its actual localization as well as to distinguish the buildings. Similarly, the boundaries between the forest and the fields look strongly enhanced with SARDINetV2.

Yet, despite the architectural modifications, SARDINetV2 is not able to reconstruct the different roads nor the smallest forest regions in the images. It illustrates the strong capabilities in translating only SAR features and thus being relevant with respect to the input as it cannot hallucinate features that are invisible in the input image.

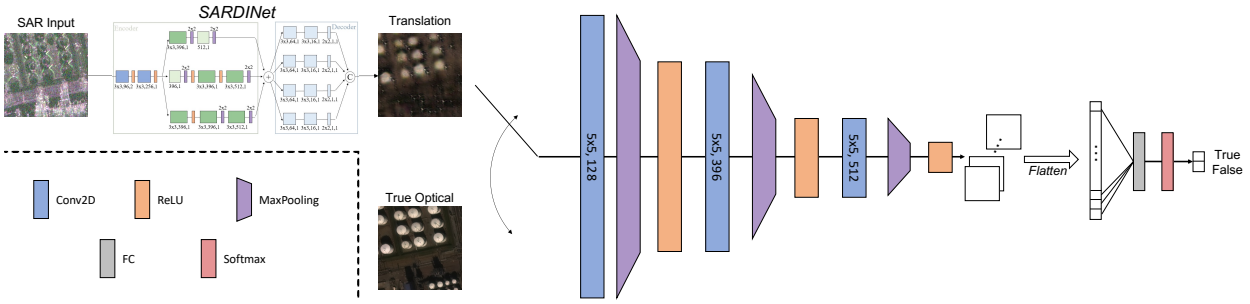


Figure 4.12: Architecture of the discriminator used for the training of Adv-SARDINet.

4.5.2 Adversarial Translations

Based on the extensive literature on modality translation in both standard computer vision and remote sensing, Generative Adversarial Networks (GAN) are usually preferred over generator-only architectures. By using an auxiliary classifier network that aims to discriminate whether the input is a true image or a generated one, the generator network is forced to produce credible reconstructions to fool the auxiliary classifier - also known as the discriminator. Because of its strong ability to generate highly credible patterns, this section explores an adversarial extension of SARDINet, which is described in [Bralet et al., 2022] and named as Adv-SARDINet.

4.5.2.1 Adv-SARDINet

To comply with standard GAN architectures, a discriminator network is implemented to evaluate the quality of SARDINet reconstructions compared to optical ground truths. The architecture given in Figure 4.12 follows a standard classifier template with three convolutional blocks consisting of Conv2D, MaxPooling and ReLU layers. A final Fully Connected (FC) layer is applied after the flattening of the last features to give the final decision.

The architecture is deliberately kept very simple to reduce the high risk of hallucinating unreliable artifacts during the training. Indeed, this shallow architecture forces the discriminator to focus only on the most discriminative features. The main drawback of such a strategy is that the discriminator may have less ability to focus on high-resolution details. To mitigate this issue, a large number of convolutional kernels is used to increase the variety of the patterns of attention.

4.5.2.2 Results

The experiments on Adv-SARDINet are performed on SpaceNet 6 [Shermeyer et al., 2020], following the same configuration as the original SARDINet architecture in [Bralet et al., 2022] and presented in Table 4.1 of Section 4.3.2. Based on the quantitative and qualitative performance of Adv-SARDINet in Table 4.9 and Figure 4.13, respectively, advantages and drawbacks can be drawn.

From a quantitative point of view, Adv-SARDINet suffers from significant performance degradation. This phenomenon may be due to the generation of bright artifacts caused by the small scatterers in the input SAR image. In addition, the smaller tanks in the second row of Figure 4.13 are not reconstructed

Network	↓ RMSE	↓ MAE	↑ SSIM-3	↑ SSIM-11
SARDINet	**0.080**	**0.053**	**0.850**	**0.813**
Adv-SARDINet	0.096	0.063	0.802	0.762

Table 4.9: Quantitative comparisons of SARDINet and Adv-SARDINet on SpaceNet 6. Best results are in bold.

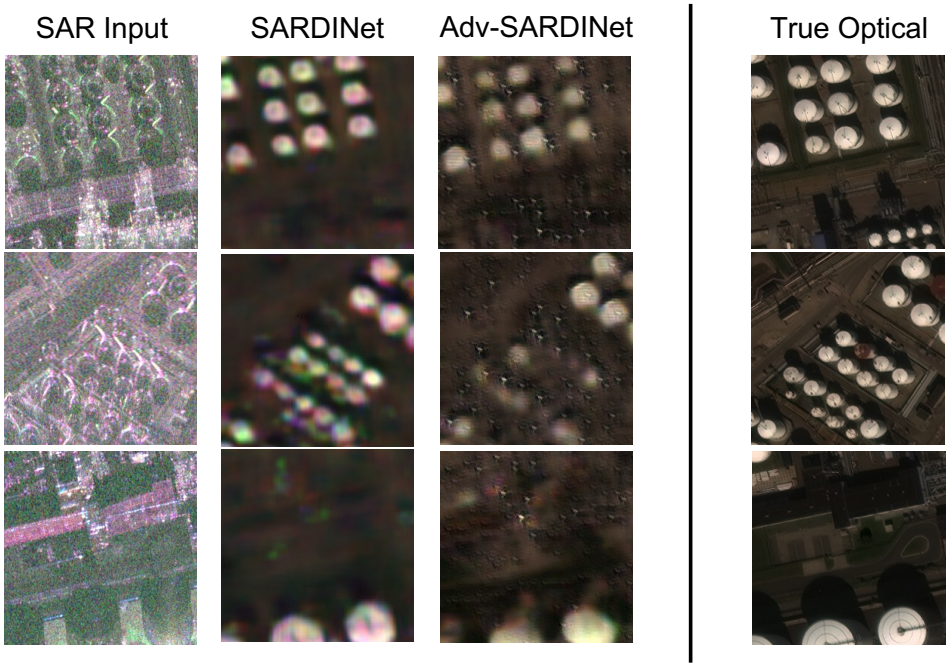


Figure 4.13: Qualitative comparisons between SARDINet and Adv-SARDINet on SpaceNet 6.

with Adv-SARDINet which further reduces their relevance. Finally, from a training point of view, Adv-SARDINet is subject to divergence after 250 epochs which prevents to train the network longer and to reach the performance obtained with SARDINet-400 in Section 4.3.2.

Yet, from a qualitative point of view and despite the artifacts, Adv-SARDINet shows two main improvements compared to SARDINet. First, the boundaries of the tanks are better reconstructed and less blurred than in the original SARDINet version as it can be seen in the first two rows. Second, Adv-SARDINet does not focus especially on the reconstruction of the tanks but achieves visually accurate results in the darkest areas of the image. In particular, in the first and third rows, it is easier to distinguish the buildings, the roads and the tank platforms than in the reconstructions of SARDINet. In addition, the artifacts generated by Adv-SARDINet are not due to inaccurate hallucinations but to the bright pixels from the input image which are misunderstood by the network. Thus, the simple architecture of the implemented discriminator successfully limits the drawbacks of the standard GAN but trades such behavior for its noise robustness.

4.5.3 Post-translation Deblurring

Since modality translation is already an ill-posed problem and therefore a challenging task for a neural network, constraining it to enhance the contrast makes the task even more difficult. Another solution to improve the contrast of the different translations lies in the use of a post-processing stage. This section, gives more details on the standard image processing methods used in [Bralet et al., 2024c] to improve the final contrast.

4.5.3.1 Standard image processing

Two standard deblurring approaches are explored in this work to make the translated images $\mathcal{I}$ more credible, namely the Laplace and the Wiener filters.

The Laplace filter: This approach consists of convolving the blurred image with a weighted version of the Laplace filter so that the boundaries of the objects can be better delineated. Mathematically speaking,

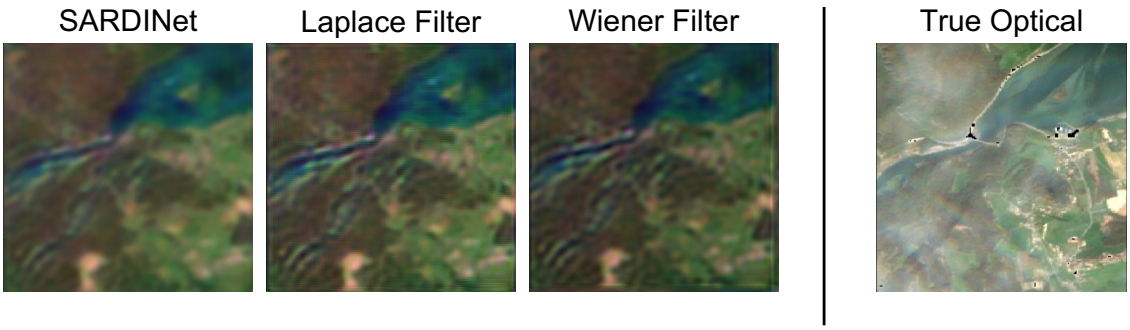


Figure 4.14: Qualitative results following the application of post-translation filters on SARDINet results.

Equation (4.1) gives the calculation of the enhanced image.

$$\mathcal{I}_{Lapl} = \mathcal{I} * \begin{bmatrix} -1 & -1 & -1 \\ -1 & 9 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{4.1}$$

The Wiener filter: This filter assumes that the image suffers from a Gaussian blur. The idea is then to divide the Fourier transform of the blurred image by the Fourier transform of a Gaussian kernel. Transforming the image back into the spatial domain results in the final enhanced image. The specificity of the Wiener filter is that a normalization factor W is applied before the inverse Fourier transform. Mathematically speaking, Equation (4.2) gives the overall process.

$$\mathcal{I}_{Wien} = IFFT(W \frac{FFT(\mathcal{I})}{FFT(G(\mu,\sigma))}) \qquad \text{with} \quad W = \frac{1}{1 + \frac{c}{|FFT(G(\mu,\sigma))|^2}} \tag{4.2}$$

where FFT and $IFFT$ are the Fourier and the inverse Fourier transforms, $G(\mu,\sigma)$ is the Gaussian kernel of mean μ and standard deviation σ, and c a constant.

4.5.3.2 Results

The filters are applied to the results of SARDINet trained on the mountain dataset with the configuration described in Section 4.4.1. The Gaussian kernel used for the Wiener filter is computed with a kernel size of 9×9, with the mean μ set to 0 and the standard deviation σ set to 29, and the constant c set to 0.8. This parameter configuration is selected because it obtains the most relevant results compared to the other tested configurations.

The improvements obtained with the two standard deblurring strategies on SARDINet are illustrated in Figure 4.14 and Table 4.10. From a qualitative point of view, the most striking point is the strong enhancement of the boundaries between the land-cover classes - *e.g.* the lake with the island, the forest

	↓ MSE $_{(.10^{-2})}$	↑ SSIM $_{(.10^{-2})}$	↑ PSNR	↓ FID
SARDINet	2.91±0.06	**74.44**±0.07	**28.58**±0.04	38.73±0.54
w. Laplace Filter	**2.83** ±0.04	71.67±0.03	28.47±0.03	**24.05**±0.90
w. Wiener Filter	3.36±0.06	72.61±0.08	28.47±0.03	31.65±0.57

Table 4.10: Quantitative impact of the application of the filters on the translations of SARDINet on the mountain dataset. Best results are displayed in bold.

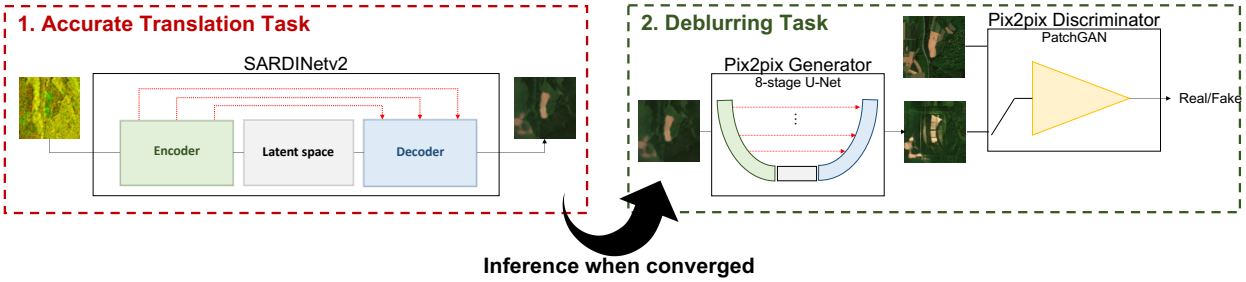


Figure 4.15: Proposed 2-step Translation-Deblurring framework. The SAR image is first translated into an optical image with SARDINetV2 architecture. Once the training ends, the latter is frozen and a Pix2pix network is trained to deblur the translated image.

with the fields and the buildings themselves. This has a direct impact on the quantitative performance since the FID score of the final image after the Laplace (resp. Wiener) filter reduces from 38.73 to 24.05 (resp. 31.65). Such a decrease in the FID score makes SARDINet results even more competitive with those of the adversarial networks, Pix2pix and CycleGAN. Indeed, they achieve a FID score of 21.71 ± 8.67 and 10.67 ± 13.14, respectively, in Table 4.4 of Section 4.4.1. More interestingly, the application of these filters does not strongly impact the relevance metrics. Indeed, as visible in Table 4.10, the filters maintain similar MSE and PSNR values while decreasing more significantly the SSIM metric which remains much better than the performance of CycleGAN and Pix2pix observed in Table 4.4.

From a credibility point of view, the main drawback of this strategy is the lack of hallucination capabilities. Indeed, although forest patterns can be guessed in the filtered images in Figure 4.14, they are still far from being as credible as those reconstructed with GANs. In addition, both filters show defects at the boundaries of the image, which reduces their applicability to patches where the phenomena of interest are in the center of the image. Nevertheless, from a relevance point of view, these simple post-processing strategies help to preserve the consistency of the information reconstructed by SARDINet while improving the sharpness of the transitions from an object to the other.

4.5.4 Translation-Deblurring Framework

Since hallucinated patterns generated by GANs are mostly creating artifacts in the final images, a wise method would be to avoid their use. Yet, the textures of underlying land-cover can also be valuable information for downstream classification tasks. Based on the partially satisfactory results obtained with the three previous methods, each of them brings advantages for the reconstruction of less blurred images. Therefore, in this section, we propose a preliminary study that combines the three previous intuitions into a Translation-Deblurring framework (Tra-De) which uses GANs as a post-processing for texture reconstruction.

4.5.4.1 Deep adversarial post-deblurring

As a reminder, the main objective is to improve the contrast of the reconstructed image while preserving the relevance of the patterns obtained when translating the SAR images with SARDINet. Therefore, the proposal is to decouple the translation and deblurring tasks as experimented in Section 4.5.3. Then, as depicted in Figure 4.15, each task integrates the improvements mentioned in Section 4.5.1 and 4.5.2.

On the one hand, SARDINetV2 is shown to increase the relevance of the patterns but also their credibility during the translation process. Therefore, by applying it to the translation task, the idea is to ensure the relevance of the reconstructed patterns and to remain consistent from a structural perspective between the SAR and the optical image. On the other hand, adversarial training have shown strong capabilities in deblurring and hallucinating high-resolution details but lowered the quantitative performance. Then, by applying a standard GAN architecture (*i.e.* Pix2pix [Isola et al., 2017]) for the deblurring task only, the

Translation Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	BigEarthNet-MM	Loss:	MAE + SSIM	Metrics:	MSE, PSNR, FID
Input:	SAR (VH, VV)	Optimizer:	Adam		
Output:	Optical (RGB)	Batch Size:	16	SOTA:	Pix2pix [Isola et al., 2017]
Size:	120×120 px	Learning Rate:	5.10^{-5}		CycleGAN [Zhu et al., 2017]
#Training:	156.087	#Epochs:	30		SARDINet [Bralet et al., 2022]
#Evaluation:	39.022				SARDINetV2 [Bralet et al., 2024b]

Specificities for Deblurring
A Pix2pix network is applied to the output of SARDINet and SARDINetV2 (RGB images).
The original loss of Pix2pix is used with a learning rate of 1.10^{-4}.
The remaining of the deblurring configuration is unchanged.

Table 4.11: Translation-Deblurring Framework training configurations. SOTA is standing for State-Of-The-Art.

idea is to help the network to focus on enriching the already accurate features with the most probable high-resolution patterns.

Both networks are trained successively and independently to ensure that SARDINetV2 achieves accurate translation performance before training a Pix2pix model on already converged translated images. Pix2pix is selected among the various approaches in the literature for two main reasons. First, it is a supervised GAN which can help to regularize the network and limit the generation of artifacts in the final image. Second, Pix2pix is an architecture that has already demonstrated strong performance in several tasks in computer vision. Indeed, despite the fact that its generator is a standard U-Net architecture with eight downsampling stages, its discriminator is especially leveraged in the literature. The PatchGAN discriminator consists of a succession of three stages of 2D convolutions and Leaky ReLU layers. Contrary to the discriminator proposed in Section 4.5.2, PatchGAN divides the input image into patches and evaluates whether each patch is real or artificially generated. The final decision is based on an average of the patch decisions. Thanks to this local strategy, PatchGAN ensures that the high-resolution patterns of the reconstructions of the generator are consistent with those of the true images.

4.5.4.2 Preliminary results

The experimental setup used to obtain the preliminary results is summarized in Table 4.11. To ensure a correct convergence and generalizability of the translators, the experiments are conducted on BigEarthNet-MM [Sumbul et al., 2021] described in Section 4.2.2. The training of the deblurring Pix2pix is then performed downstream to the translations of SARDINet and SARDINetV2 with the same hyperparameters as the translation networks except for the learning rate which is set to 1.10^{-4}. To be consistent with the previous experiments and to evaluate the benefits of the proposed Tra-De framework, the MSE and PSNR are used to evaluate the relevance of the reconstructed patterns while the FID score gives an idea of their credibility with respect to the true optical images.

Table 4.12 and Figure 4.16 describe the quantitative and qualitative results obtained by the translation networks compared to those obtained according to the Tra-De framework, respectively. As expected in this study, standard adversarial networks achieve the best FID score and SARDINet-based architectures achieve the best relevance metrics. Nevertheless, both of these approaches fail to reconstruct relevant and credible results. By implementing the Tra-De framework on the results of SARDINet and SARDINetV2, the metrics are much better balanced than on the original trainings. Indeed, the FID score is reduced by more than 6 and 5 points for SARDINet and SARDINetV2, respectively. But this strong increase in credibility has a cost on the relevance metrics as the MSE score increases by $0.33.10^{-3}$ for SARDINetV2 and its PSNR decreases by more than 0.7 points. Nevertheless, comparing such performance with that of the standard adversarial methods shows the actual benefits of the Tra-De framework since the latter still outperforms

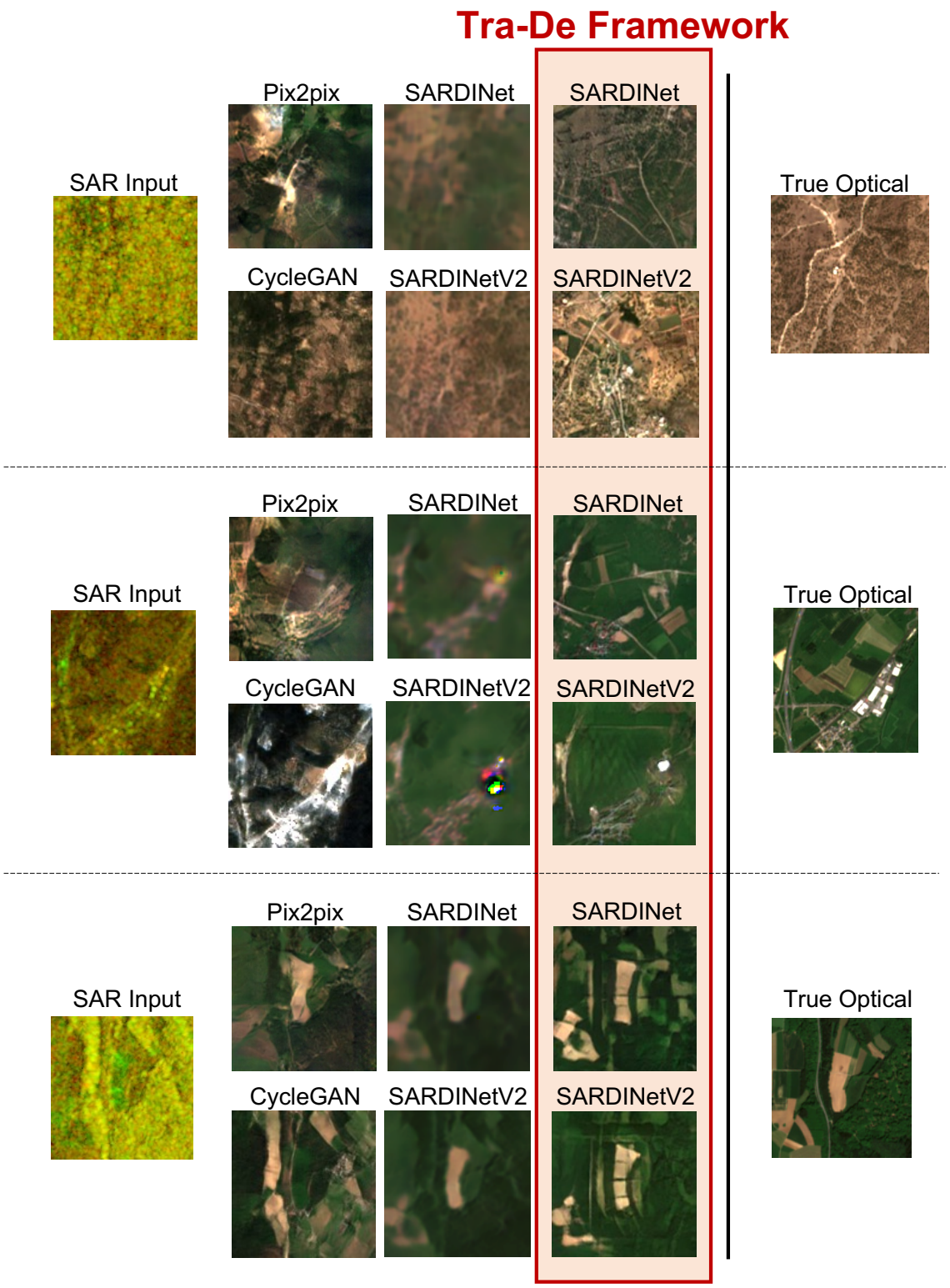


Figure 4.16: Qualitative results obtained with translation-only networks and after applying the Tra-De framework on SARDINet-based translators (red box).

	PSNR ↑	MSE ↓	FID ↓
Pix2pix [Isola et al., 2017]	21.93	$18.48.10^{-3}$	2.67
CycleGAN [Zhu et al., 2017]	23.25	$26.50.10^{-3}$	0.54
SARDINet [Bralet et al., 2022]	25.34	$11.19.10^{-3}$	7.46
SARDINetV2 [Bralet et al., 2024b]	25.47	$10.19.10^{-3}$	6.32
Tra-De w. SARDINet	24.65	$10.62.10^{-3}$	0.60
Tra-De w. SARDINetV2	24.73	$10.52.10^{-3}$	0.55

Table 4.12: Quantitative performance of the Tra-De framework compared to standard translation methods. In the table, the greener the better, the redder the worst.

Pix2pix by 8.10^{-3} points in MSE score and CycleGAN by 1.48 points in PSNR. Therefore the trade-off made by the Tra-De framework to achieve more credible results against their relevance looks still acceptable. In particular, the performance of the Tra-De framework applied to SARDINetV2 obtains the best balance in between the credibility and relevance metrics with the second best FID and MSE scores and the third best PSNR score.

The observed decrease in the relevance performance can be explained by the qualitative results in Figure 4.16. Indeed, since the PSNR is related to the measure of the apparent noise in the image, it is extremely sensitive to the generation of artifacts in the images. Since the Tra-De framework is based on an adversarial architecture (Pix2pix), hallucinated artifacts are generated in the final images such as the dense forest or the urban patterns in the first row for SARDINet and SARDINetV2, respectively, or the creation of roads in the middle of the fields in the second row. Nevertheless, the images are much more credible compared to their translation-only counterparts and more accurate compared to their adversarial competitors. Indeed, the sparse forest of the first row which is successfully reconstructed by SARDINetV2 is still visible in the results of the Tra-De framework applied on SARDINetV2, the distinction between the forest and the fields in the third row is also maintained and accurately enriched with forest patterns. Then, compared to the adversarial networks, the city in the second row is better discriminated with Tra-De trained on SARDINetV2 than on CycleGAN - or even on Pix2pix where the city is confused with field patterns.

Therefore, based on these preliminary results, the performance of the Tra-De framework seems to achieve a better relevance-credibility balance than the translation-only networks whether adversarial or SARDINet-based. Further experiments should be performed to evaluate the stability of the results and their applicability to a downstream task. In particular, the trade off made by the Tra-De framework by generating hallucinations and thus reducing the relevance of the patterns can be severe enough to prevent its further use for downstream tasks, making it only interesting for the visualization of attractive and credible images. Yet, this study is encouraging and shows that simplifying the tasks of the networks helps to meet both the quantitative and qualitative expectations regarding modality translation.

4.5.5 Contrast Enhancement Outcomes

Following the various experiments proposed in this section, and despite the early stage of the final Translation-Deblurring framework, the relevance of SARDINet for SAR-to-optical translation can be settled. Indeed, the simple architectural modifications leading to SARDINetV2 in Section 4.5.1 demonstrate the improvement possibilities of the original SARDINet but also its great stability in SAR distortion robustness. In addition, the application of post-translation deblurring techniques in Section 4.5.3 successfully preserves the relevance of the patterns while significantly improving the discrimination from one class to another. Nevertheless, these approaches cannot reconstruct the textures that are characteristic of the underlying land-cover classes. An adversarial approach is implemented in Section 4.5.2 and shows strong contrast enhancement capabilities but at the cost of pattern relevance and training instability. Therefore, to take the best of the three worlds, a Translation-Deblurring framework is designed in Section 4.5.4 and shows encouraging preliminary results, including a strong contrast enhancement and a correct transfer of the relevance of the patterns reconstructed by SARDINetV2.

4.6 Conclusion

In this chapter, modality translation is introduced as a key step for better understanding of SAR images and is intended to be used to improve the quality of slope instability detection after their collapse has occurred. To this end, a novel SAR Distorted Image translator Network (SARDINet) is proposed to perform SAR-to-optical translation of SAR images suffering from geometrical distortions. Following specific focuses on the fusion and decoding strategies of the network, SARDINet demonstrates strong capabilities to address the reconstruction of optical images in distorted urban and mountainous areas. Nevertheless, the lack of contrast in the translated image makes it less credible from a qualitative point of view. Therefore, other ways to enhance the contrast of the final optical image are explored through architectural modifications (SARDINetV2), adversarial training (Adv-SARDINet) or post-translation image processing. The Translation-Deblurring framework (Tra-De) emerges from these three studies and, according to preliminary results, achieves the best quantitative-qualitative balance. Nevertheless, the latter is still at an early stage of study and the quantitative metrics are lower compared to the original SARDINet and SARDINetV2 architectures. Since modality translation is intended to be used for downstream tasks, the relevance of the patterns is favored over their credibility and the decision is made to keep the standard SARDINetV2 architecture in the following chapter.

Related communications

Conference

[Bralet *et al.*, ICIP 2022] *Deep Learning of Radiometrical and Geometrical Sar Distorsions for Image Modality translations*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E. (2022, October). In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 1766-1770). IEEE. DOI: 10.1109/ICIP46576.2022.9897713.

[Bralet *et al.*, GRETSI 2023] *Impact de la stratégie de décodage sur la traduction de modalité radar-optique d'images de télédétection*, Bralet, A., Atto, A., Chanussot, J., Trouvé, E., number 2023-1309, pages p. 929–932, Grenoble. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

[Bralet *et al.*, IGARSS 2024] *DEM-assisted neural network for SAR-to-optical image translation*, Bralet, A., Ngo, T. N., Trouvé, E., Chanussot, J., Atto, A. M., July 2024, submitted to IEEE International Geoscience and Remote Sensing Symposium (IGARSS) 2024.

Journal

[Bralet *et al.*, CVIU 2024] *Translation-classification loss for SAR image understanding with deep learning*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E., submitted to Computer Vision and Image Understanding (CVIU), Elsevier - submitted April 19, 2024

Dataset

[Bralet *et al.*, IEEE Dataport 2024] *SAR-DEM-optical mountainous dataset for distortion management*, Bralet, A., Ngo, T. N., Trouvé, E., Chanussot, J., Atto, A. M., January 2024, IEEE Dataport, doi: https://dx.doi.org/10.21227/16ce-ce92

Chapter 5

Sudden Landslides Detection

5.1 Introduction

The accurate translations proposed in the previous chapter strongly encourage their use for performing downstream tasks. Since this thesis focuses on the detection of slope instabilities, we propose to use the translations in a change detection context. Due to the fact that slow moving slope instabilities cannot be properly addressed by either optical images or SAR images without InSAR processing (see Chapter 6), this chapter focuses on post-collapse landslide detection. In this chapter, translations are used as a way to overcome the cloud cover that can be caused by intense rainfall events that trigger landslide collapse. Therefore, only post-event images are translated in this work, since a pre-event image is supposed to be available and only landslide-related changes are targeted. Throughout the chapter, we propose different strategies to improve the translation capabilities of SARDINetV2 with respect to the landslide change detection task on a newly proposed challenging dataset. SARDINetV2 is chosen among the different implemented SARDINet versions since it has been demonstrated as the most relevant network in Chapter 4.

The novel multimodal bitemporal landslide dataset adapted for change detection with deep learning approaches is first described in Section 5.2. The change detection applied to the translated images resulting from the fine-tuning of SARDINetV2 on this particular dataset is then proposed in Section 5.3. The detection performance is compared to the monomodal optical and monomodal SAR change detection performance to evaluate the relevance of the proposed method on a real case study with respect to standard configurations. A novel Translation-Classification Loss (TCL) is then proposed in Section 5.4 to improve the ability of the translator to reconstruct more accurate land-cover patterns. This TCL is finally extended into a Translation Change Detection Loss (TCDL) to specialize the network in reconstructing patterns specific to the landslide change detection task.

5.2 Materials

Two datasets are of particular interest in this chapter. First, the BigEarthNet-MM dataset [Sumbul et al., 2021] described in Section 4.2.2 is exploited as a pre-training dataset. Due to the variety of landscapes, the number of SAR-optical pairs available, and the multi-label annotations provided by the authors, this dataset is particularly rich. Second, a smaller dataset is proposed to focus on the task of interest: post-collapse landslide detection. Despite the availability of the Landslide4Sense dataset [Ghorbanzadeh et al., 2022b], it does not fit the actual targeted task. Although Landslide4Sense is a large database of landslides, it consists only of post-event optical images. Therefore, such a dataset can be used for inventory purposes, but not for a proper change detection task since pre-event images are missing. In addition, the authors ensure that landslides are visible in each of the patches. However, in practice, early detection of landslide collapse often means the presence of dense cloud cover, which makes the inventory optically infeasible and necessitates the use of SAR imagery, *i.e.* multimodal approaches. Due to these limitations, a new multimodal change

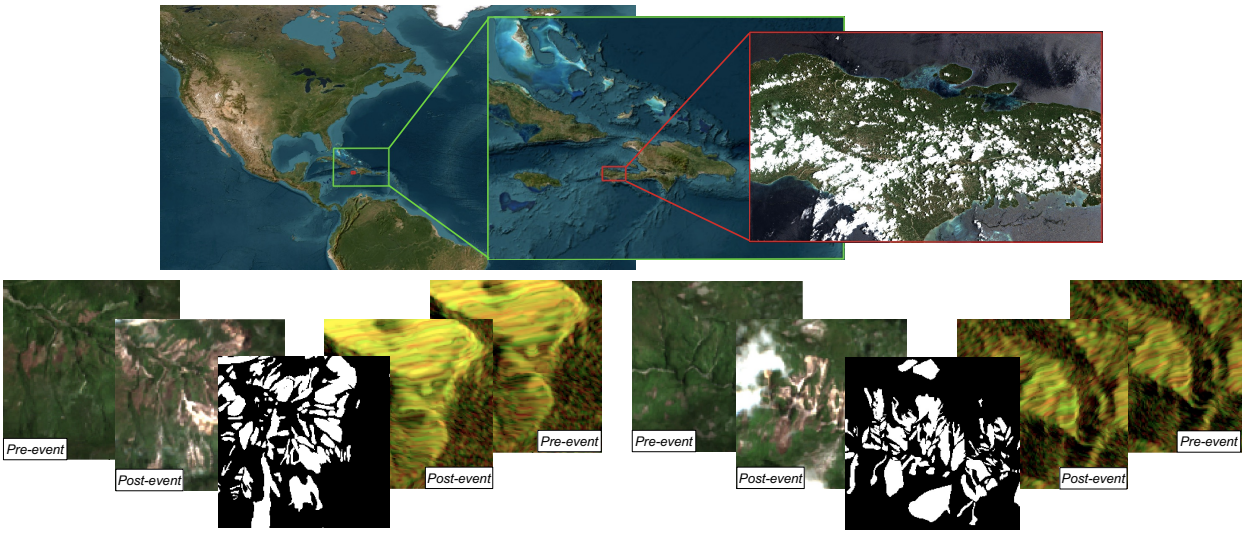


Figure 5.1: Location of the proposed dataset in Haiti (top) with the post-earthquake image of Sentinel-2 acquired on August 14, 2021. At the bottom, two samples of the dataset with pre- and post-event optical images acquired on August 04, 2021 and August 14, 2021, and their corresponding SAR images in the ascending passes on August 05, 2021 and August 17, 2021, respectively. The landslide annotation maps are displayed in the center of each sample.

detection dataset for sudden landslide detection is proposed in this thesis.

The dataset is based on Sentinel-1 and Sentinel-2 data acquired before and after the magnitude 7.2 earthquake that struck Haiti on August 14, 2021. After the event, the National Aeronautic Space Administration (NASA) provided an inventory of triggered landslides[1] containing $14,482$ polygons. Among these annotations, a total of $11,993$ landslides are identified with an area smaller than $5,000\text{m}^2$, $1,548$ landslides have an extent between $5,000\text{m}^2$ and $10,000\text{m}^2$, and 941 landslides are larger than $10,000\text{m}^2$. Figure 5.1 illustrates the location of the annotated landslides in Haiti. To avoid revegetation over the landslides which can reduce their detectability, but also to mimic the inference time process, the SAR and optical acquisitions are chosen as temporally close to the event as possible. On the one hand, the Sentinel-1 images are GRD products acquired in IW mode in both ascending and descending passes. The ascending (resp. descending) pre- and post-event images have been acquired on August 05, 2021 (resp. August 03, 2021) and August 17, 2021 (resp. August 15, 2021). A standard pre-processing routine similar to BigEarthNet-MM [Sumbul et al., 2021] is applied to each image. It consists of the orbital path correction, a termal and border noise removal and σ_0 calibration. The resulting images are then orthorectified with the SRTM 1sec HGT DEM [Farr et al., 2007] to match the optical image georeferencing, and transformed to a logarithmic scale to reduce the long tail distribution of the values. On the other hand, the pre- and post-event Sentinel-2 Level-2A optical images are acquired on August 04, 2021 and August 14, 2021. Fortunately, the cloud cover on the D-Day is still acceptable, as visible in the red rectangle of Figure 5.1. This makes this study case particularly interesting for accurate SAR-optical pairing. To better guide the identification of landslides from these images, two additional pieces of information are provided in the dataset. First, the layover and radar shadow masks computed according to the algorithm described in [Meier et al., 1993] are aggregated and provided for the ascending and descending passes. Due to the lack of SAR information in such areas, landslide detection would be more complex. The mask can be used to weight the loss function or to propose more sophisticated strategies such as curriculum learning [Bengio et al., 2009]. Second, a cloud-related mask is also provided by merging the medium and high probability cloud masks together with the cloud shadow mask computed during the scene classification step of the Sen2Cor processor[2].

[1] https://maps.nccs.nasa.gov/arcgis/apps/MapAndAppGallery/index.html?appid=574f26408683485799d02e857e5d9521, last visited on July 15, 2024.

[2] https://step.esa.int/main/snap-supported-plugins/sen2cor/, last visited on March 17, 2024

To make the dataset available for deep learning approaches, these images are cropped around the location of the landslides. After rasterizing the landslide shapefile, each image is coregistered with respect to the orthorectified SAR image acquired on August 03, 2021 to ensure the correspondence of each pixel from one image to another. The patch extraction process is based on a sliding window strategy with a size of 128×128 to match the same dimensions as BigEarthNet-MM and with a stride of 64 in both spatial dimensions to multiply by 4 the number of images in the dataset. The sliding window aims to identify all patches containing more than 100 pixels labeled "landslide". The patches that meet this condition are then cropped on each of the images and masks. The extraction process results in $1,713$ patches for each image, two examples of which are given at the bottom of Figure 5.1. This allows the dataset to be used under different configurations:

- Optical-optical change detection: Among the $1,713$ patches extracted, only 205 patches have less than 10% cloud cover in both the pre- and post-event images. Therefore, the cloud-free dataset is particularly small which is why all images are provided to exploit the cloudy images as weak samples for the training.
- SAR-SAR change detection: Since SAR images are weather independent and acquired in both ascending and descending passes, all ascending-ascending and descending-descending patch pairs can form the SAR-SAR dataset. Such a dataset would therefore contain a total of 3,426 fully exploitable samples associated with their respective geometrical distortion masks.
- Optical-SAR change detection: This configuration is expected to be the most interesting at inference time since the SAR image is necessarily exploitable after the event. By merging ascending and descending images, the dataset can also be composed of $3,426$ images.

The proposed dataset is made available on IEEE Dataport[3] to encourage the development of sudden landslide change detection approaches based on multimodal data.

5.3 Translation-based Landslide Detection

The objective of this section is to explore the capabilities of a standard change detection neural network in finding slope instability failures following the earthquake that occurred in Haiti on August 14, 2024. Based on the dataset constructed in Section 5.2, the task is particularly complex. First, the number of non-cloudy optical pairs is very limited which makes the training prone to overfitting quickly. Second, the region is mainly forested which makes the SAR acquisitions very noisy and difficult to interpret. Finally, the hilly nature of the region causes strong geometrical distortions in the SAR images, which limits the observation of landslides in these particular areas.

5.3.1 Training Strategies

To investigate the applicability of modality translation in the specific context of Haitian landslides, several experiments are conducted. To fully exploit the training dataset, two mono-modal change detection strategies are first explored. These two configurations are considered as references to compare the performance of a third strategy consisting of applying the change detection to a pre-optical image and a post-translated image.

Optical-optical: Despite the presence of clouds in both the pre- and post-event images, optical-optical change detection remains the most reliable strategy. Based on the radiometrical changes induced by the landslides, these phenomena can be easily identified when comparing the two images, as visible in Figure 5.1. Therefore, this experiment is expected to achieve the best performance that the optical-translation configuration is intended to approach.

SAR-SAR: Since SAR images are weather independent, they can be acquired despite the presence of dense cloud cover. Therefore, this strategy is always applicable to the detection of changes caused by natural hazards. Nevertheless, landslides are much more difficult to identify on such images (see Figure 5.1). This

[3]https://ieee-dataport.org/documents/multimodal-remote-sensing-dataset-landslide-change-detection-haiti, last visited on July 16, 2024.

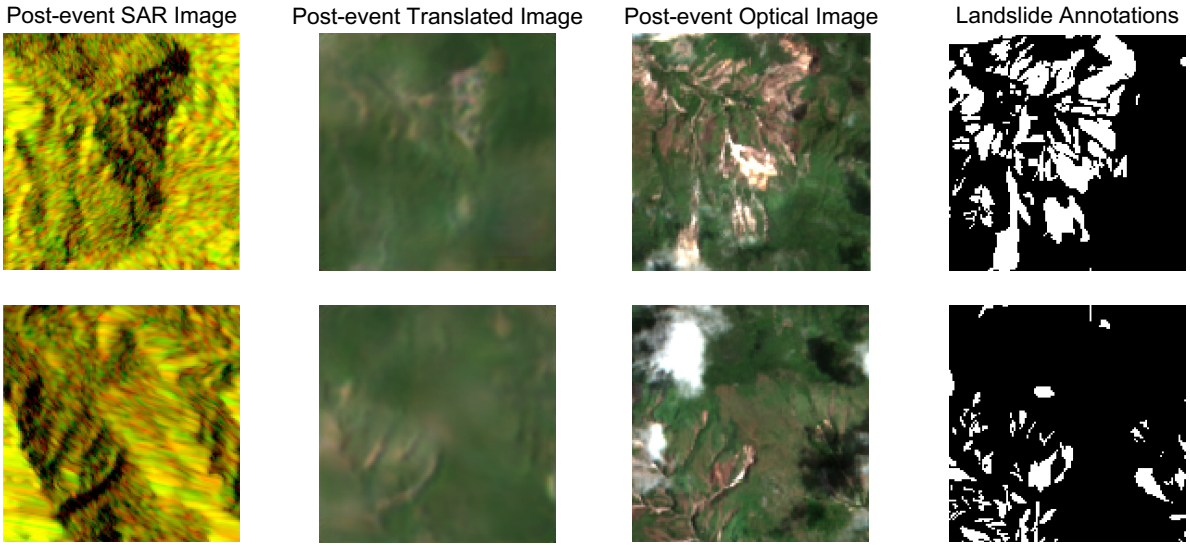


Figure 5.2: SAR-to-optical translation results of SARDINetV2 fine-tuned on the Haiti dataset

configuration acts as a lower bound in terms of performance since it is the all-weather reference method. Thus, the proposed optical-translation configuration is intended to overcome SAR-SAR change detection.

Optical-translation: This strategy aims to leverage the all-weather capabilities of SAR images and the radiometrical power of optical images. It consists of translating the post-event SAR image into an optical image and then applying a change detection deep network to the optical-translation pair. The main goal of this configuration is to outperform the SAR-SAR strategy and achieve insightful segmentation prior to the acquisition of an exploitable post-event optical image.

5.3.2 SARDINetV2 Fine-tuning

Since SARDINetV2 is shown in Chapter 4 to be the most efficient network in terms of relevance and credibility, it is chosen for the translation of the SAR images acquired after the landslide collapse. To limit the possible overfitting due to the small size of the dataset described in Section 5.2, the parameters of the networks are initialized with those resulting from the experiments on BigEarthNet-MM [Sumbul et al., 2021] in Section 4.5.1. The network is fine-tuned for 40 epochs with a learning rate of 1.10^{-6}. The Adam optimizer updates the parameters of the network based on the sum of the MAE and SSIM used to compare the translated image $\hat{\mathcal{I}}$ and its corresponding true optical image $\mathcal{I}^{Opt}$. This loss is weighted according to the landslide annotations A^L as given in Equation (5.1) to limit the effect of the class imbalance associated with the larger number of non-landslide pixels.

$$Loss(\hat{\mathcal{I}}, \mathcal{I}^{Opt}, A^L) = \sum_{(i,j)\in H\times W} (MAE(\hat{\mathcal{I}}_{ij}, \mathcal{I}^{Opt}_{ij}) + SSIM(\hat{\mathcal{I}_{ij}}, \mathcal{I}^{Opt}_{ij})) * (1 + \alpha A^L_{ij}) \tag{5.1}$$

where α is a hyperparameter used to force the network to focus on reconstructing accurate landslide patterns. After several experiments with α varying from 1 to 50, its value is fixed at 20 for fine-tuning SARDINetV2. This allows to obtain the performance shown in Figure 5.2. Despite the blurry aspect of the translations, a certain number of landslides can be identified more easily than in the SAR images which makes the translations promising for the subsequent change detection task.

5.3.3 Change Detection Architectures

Similar to the fusion study in Section 4.4.1, where the SAR image is fused to the DEM to improve the consistency of the modality translation task, this section explores three different fusion strategies for change

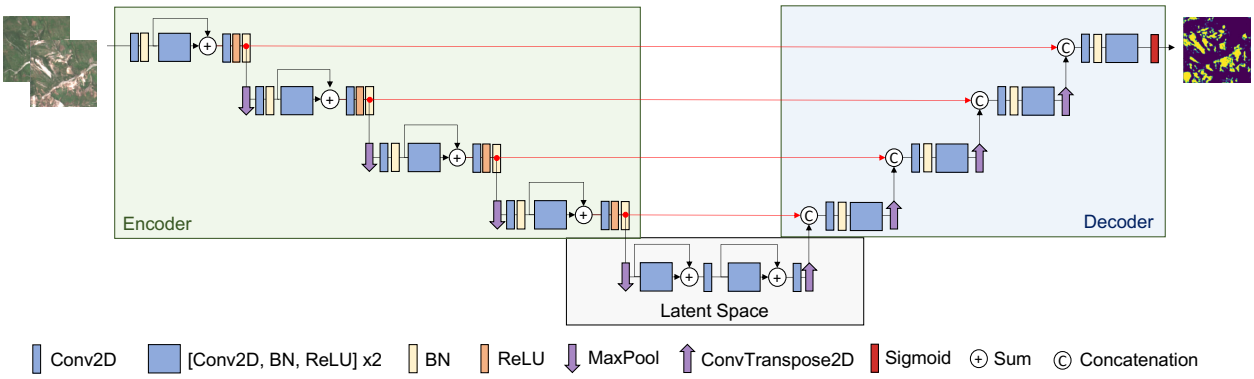


Figure 5.3: Architecture of the backbone CDNet.

detection. The two standard early and late fusions are necessarily considered to evaluate whether the images should be processed together or separately. The third approach is the Siamese strategy [Bromley et al., 1993]. Indeed, Siamese networks have already shown high performance for change detection on images [Zhang et al., 2018c, Ebel et al., 2021, Amankwah et al., 2022] which explains their use in this section.

5.3.3.1 Network architecture

The network used for change detection, called CDNet, is based on a four-stage U-Net [Ronneberger et al., 2015] as visible in Figure 5.3. Each encoding stage consists of a series of Conv2D, BN, residual block and ReLU to extract the most relevant features for each resolution. The implemented residual blocks contain two successive series of Conv2D, ReLU, BN before adding the residual connection and applying a final Conv2D layer to smooth the features. Each encoding stage ends with a MaxPooling layer that transfers the features to the lower resolution stage. The resulting latent code is further processed by two residual blocks before being fed to the decoder. This decoder is made symmetrical to the encoder to better integrate the high-resolution features extracted by the encoder. Indeed, the features preceding each MaxPooling layer in the encoder and marked with a red dot in Figure 5.3 are transferred to the corresponding decoder stage via skip connections. These features are concatenated with the corresponding resolution features resulting from a transposed convolutional layer in the decoder. The final change detection map is processed by a Sigmoid activation layer to ensure that each pixel belongs to $[0, 1]$.

The architecture of the backbone network is deliberately kept simple since the objective is to evaluate the applicability of SAR-to-optical translation for change detection. More than a network architecture, it is the concept, its relevance, advantages and drawbacks that are evaluated in this section.

5.3.3.2 Fusion strategies

Although the three fusion strategies are based on the same U-Net architecture, some modifications should be considered. The early fusion change detection network (CDNet_E) concatenates both the pre- and post-images along the channel dimension. The resulting image is then fed directly into the network to generate the change detection map. Conversely, the late fusion change detection network (CDNet_L) requires two independent encoders. Each of the pre- and post-event images passes through its dedicated encoder. The latent codes are then concatenated along the channel dimension before entering the latent layers. Similarly, the features of each encoder used by the skip connections (marked with red dots in Figure 5.3) are concatenated together before being transferred to the decoder. Finally, the Siamese change detection network (CDNet_S) is based on a single encoder. Yet, each image is processed independently by this same encoder, which is also known as the weight sharing strategy. Then, instead of concatenating the extracted red dotted features, their absolute difference map is computed and transferred to the decoder through the latent space and skip connections. Such a strategy encourages the network to better identify the changes between the two images since the absolute difference operation deactivates the features that are similar in both images.

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	Haiti dataset	Loss♠:	BCE Loss	Metrics♠:	mIoU, F1-Score, Precision,
Input:	Optical-Translation pairs*	Optimizer:	Adam		Recall, Accuracy
Output:	Change Detection Map	Batch Size:	16	SOTA:	Optical-Optical CDNet
Size:	128×128	Learning Rate:	5.10^{-5}		SAR-SAR CDNet
#Training♣:	5480	#Epochs:	Adapted**		
#Evaluation:	343				
Specificities					
* For the training, pairs of pre-/post-event are made but also pre-/pre-event and post-/post-event pairs associated with a change detection map full of 0.					
♣ The number of images may vary depending on the datatype.					
♠ The loss and metrics are computed on the masked annotations (cloud mask, layover and shadow mask). Only metrics are also computed on the raw annotations. The threshold used for the binarization is selected to maximize the F1-Score.					
** Each network is trained until first signs of overfitting appear.					

Table 5.1: Training configuration for the different CDNet variations.

5.3.4 Landslides Detection Results

5.3.4.1 Experimental setup

The training configuration of the CDNets trained with an optical-translated pair is given in Table 5.1. There are a few peculiarities to be mentioned in this work. Since the CDNet applied to optical-optical, SAR-SAR, and optical-translated images do not use exactly the same data, the number of samples of each dataset may vary. The optical-optical dataset is necessarily smaller because the images are acquired only once. Since SAR images are acquired in both ascending and descending modes, the SAR-SAR and optical-translated datasets contain twice the number of samples. In addition, to increase the number of training samples, pre-pre and post-post event pairs associated with a zero map are used for the training. The idea is to force the network to compare the two images during the training process and to avoid treating them separately in order to better focus on the changes, especially those related to landslides. The number of training images referenced in Table 5.1 is the actual number of training samples used for the optical-translation configuration. Conversely, the evaluation is only applied to pre-post event image pairs.

To avoid training the network on unreliable image pairs, but also to ensure the use of the entire Haiti dataset, the annotation maps used to compute the BCE loss for the training of each network are masked according to the constraints of each data type. Thus, the pre-event and post-event cloud masks are used for optical-optical annotations. Similarly, the pre-event cloud mask is applied to the optical-translation annotations. The layover and shadow mask is then applied to SAR-SAR annotations to limit the impact of geometrical distortions on the training. For the sake of fairness, all metrics are computed on the masked versions of the annotations but also on the raw annotations to better evaluate the pros and cons of using each data type. The binarization of the predicted map produced by the network and used to compute the metrics is chosen to maximize the F1-Score.

The number of epochs used to train each network is different for each network because of overfitting issues. Indeed, during the training, the networks can be prone to rapid overfitting due to the small size of the dataset and the difficulty in finding global landslide-related features, especially in SAR images where the backscattered signal is more complex to interpret than in optical images. Therefore, the network was stopped before the first signs of overfitting, *i.e.* when the validation loss starts to increase, to ensure comparisons as fair as possible. The number of epochs for each network is given in Table 5.2. All networks have been trained with a NVIDIA A6000 GPU and took 6 to 20 minutes to perform 10 epochs.

Network	mIoU ↑	F1-S ↑	Prec. ↑	Rec. ↑	Acc. ↑	Thresh.	#Epochs
Optical-Optical Change Detection							
$CDNet_E$	**0.384** / **0.276**	**0.545** / **0.420**	**55.11** / **55.10**	56.34 / 47.80	**94.73** / 84.74	0.03 / 0.01	34
$CDNet_L$	0.367 / 0.273	0.527 / 0.416	52.33 / 53.85	56.46 / 36.76	94.44 / **89.84**	0.02 / 0.01	34
$CDNet_S$	0.332 / 0.247	0.489 / 0.386	47.26 / 43.95	53.16 / 39.04	93.73 / 88.01	0.15 / 0.07	37
SAR-SAR Change Detection							
$CDNet_E$	0.258 / 0.251	0.404 / 0.396	34.08 / 34.56	51.32 / 47.80	87.18 / 84.74	0.19 / 0.05	10
$CDNet_L$	0.222 / 0.224	0.359 / 0.362	31.17 / 32.01	43.90 / 43.32	86.92 / 84.15	0.03 / 0.02	16
$CDNet_S$	0.192 / 0.196	0.317 / 0.322	24.19 / 25.14	49.53 / 48.08	81.94 / 78.69	0.09 / 0.08	10
Optical-Translation Change Detection							
$CDNet_E$	0.200 / 0.202	0.330 / 0.332	25.76 / 24.19	54.05 / 57.93	77.35 / 75.40	0.07 / 0.06	19
$CDNet_L$	0.196 / 0.197	0.323 / 0.325	23.22 / 23.55	**58.77** / **57.99**	74.70 / 74.58	0.04 / 0.04	13

Table 5.2: Quantitative performance of change detection networks on the validation dataset of Haiti. Precision, recall and accuracy are given in %. To better understand the performance of each configuration, the quantitative metrics are performed with respect to a masked ground truth (clouds for optical and distortions for SAR) and the actual ground truth. This is given in the Table as "masked / actual performance".

5.3.4.2 Change detection performance

The performance of the different implementations of CDNet is given quantitatively in Table 5.2 and qualitatively in Figure 5.4. Special attention should be paid to Table 5.2 where each implementation is associated with two different scores. The left value refers to the performance of the network when the landslide annotation map is masked which is done to better assess the performance of the network on the areas it can actually detect. The second number refers to the performance of the network on the actual landslide annotation map without any mask to evaluate the performance losses associated with the data constraints. Based on this performance, an obvious conclusion can be drawn: the direct application of CDNet on the translated images does not show the expected performance. However, these results also reveal the reasons for such a failure.

As expected from the study, the application of monomodal optical change detection gives the more accurate performance, especially when compared to the masked annotations (left values in Table 5.2). Due to the huge differences between a forest and a landslide pattern, the detections are necessarily easier with this modality. However, it is worth noting that the current study addresses the specific case where post-event optical images are useless due to the presence of dense cloud cover. This requires the use of a different approach for rapid landslide detection. Therefore, the following discussion focuses mainly on the comparison of the monomodal SAR CDNets with the proposed optical-translation detections.

From a quantitative point of view, with a score as low as 0.2 for mIoU and 0.33 for F1-Score, the performance of the CDNets applied to the translations is not even competitive with the performance obtained by the CDNets on the monomodal SAR pairs, *i.e.* 0.258 and 0.404, respectively. Only the recall shows a different tendency from the other metrics. Nevertheless, such a high recall value can be easily explained by looking at the qualitative performance in Figure 5.4. Both $CDNet_E$ and $CDNet_L$ trained with the translations show a clear overestimation of the landslides, highlighted by red regions in the images. Since recall does not take into account the presence of false positives, its value is necessarily high, but conversely, the precision is negatively affected by such overestimations as depicted in Table 5.2. However, this may not only be due to the use of translated images in the pipeline. Indeed, based on the monomodal SAR performance of the three CDNets, a similar behavior can be observed by comparing the recall and the precision. Rather than overestimating, these networks show a more noisy qualitative performance in Figure 5.4.

Pushing the interpretation further by considering the metrics computed on non-masked annotations gives more interest in the use of SAR images and even, to some extent, of translated images. Indeed, the comparison between masked and non-masked metrics reveals a significant performance drop for the monomodal optical CDNets reaching more than 0.1 points of loss in terms of mIoU and F1-Score and 10 points of decrease in terms of recall and accuracy. Only the precision remains stable which is legitimate since it does not take into account the misdetections (or false negatives) in its calculation. Conversely, the CDNets trained on

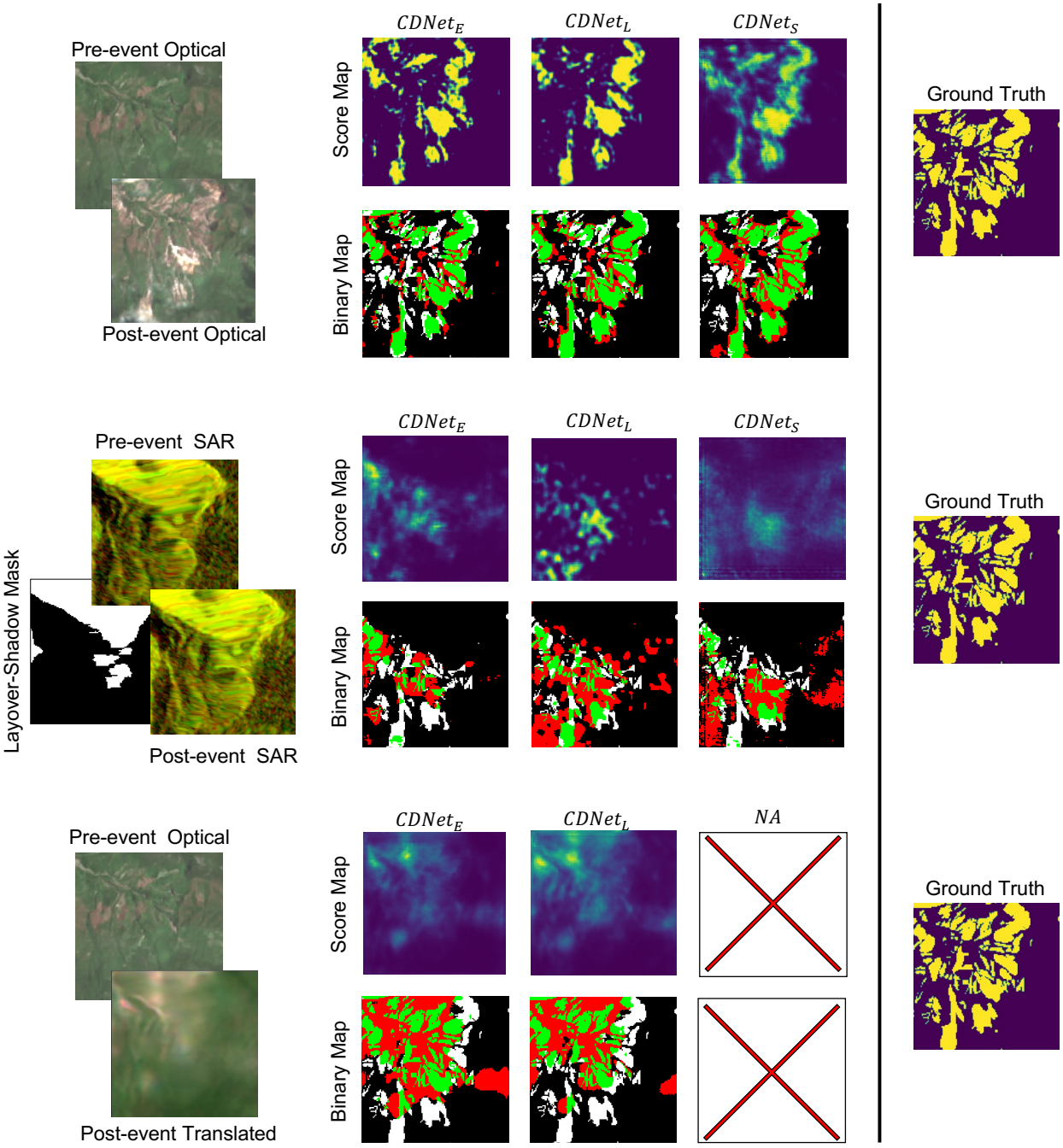


Figure 5.4: Qualitative performance of the different CDNet depending on the input data type. True positives, false positives, true negatives and false negatives are displayed in green, red, black and white on the binary maps, respectively.

monomodal SAR pairs and optical-translated pairs are almost unaffected in terms of metric values. Two reasons may explain this performance. First, due to the overestimation, it seems clear that the networks are more susceptible to detect landslides outside their effective regions. Second, although layover areas consists of an aggregation of multiple backscattered signals, there is still a signal that can be exploited (unlike shadow areas). Therefore, the network might be able to detect changes in these aggregated signals and thus highlight susceptible changes. This hypothesis remains to be proven and due to the presence of strong overestimation issues, the latter is the most susceptible reason for maintaining performance.

It is also important to note that the application of CDNet_S to the optical-translation pairs converged to a zero change map for all images within a few epochs. Due to this particular behavior, the performance is not mentioned in Table 5.2 and Figure 5.4. Nevertheless, such results illustrate a critical point regarding the translated images: the reconstructed patterns are not similar enough to the true optical patterns in terms of distribution to be processed and compared by the same encoder. Indeed, since class 0, *i.e.* no landslide, is mainly present in the images, the encoder first looks for similar patterns between the pre-event and post-event images. However, due to the fact that the images are not similar enough even in non-changed areas, the encoder cannot identify discriminative features to identify changed areas. Therefore, in the absence of discrimination and with the "no landslide" class in majority in the training dataset, the best way to minimize the loss is to output a zero change map.

5.3.5 Translation-based Change Detection Outcomes

Based on the experiments conducted in this section, the use of translations for landslide detection shows inaccurate performance. Several reasons can explain these results.

The dataset: Since SARDINetV2 has shown strong performance when trained on BigEarthNet-MM, one may question its transferability to the Haiti dataset. The latter is extremely challenging because it contains a small number of samples, it is located in highly distorted mountainous regions, and the land-cover is mainly forest, making the reconstruction of non-forest patterns even more difficult. This limited variety of patterns can affect both translation and change detection performance. In addition, the landslides triggered by the Haiti earthquake event do not backscatter strong signals compared to pre-event images, making their detection on SAR and SAR-derived images even more challenging. Finally, a large number of landslides in the Haiti dataset are very small ($< 5{,}000$ m^2). This makes them even more difficult to detect at the resolution of the Sentinel satellites and under the forest canopy.

The CDNet architecture: The architecture of the change detection networks is kept very simple in these experiments despite the challenging task addressed. Even the performance obtained on the monomodal optical pairs is not outstanding compared to standard change detection neural networks. Therefore, more attention could be paid to this particular architecture by integrating attention modules to force the network to accurately compare the pre- and post-event images, or by designing a more complex architecture with more layers, kernels and even branches. Yet, in order to keep the computational cost reasonable, this research direction should be carefully addressed and optimized.

The translation process: Due to the pre-training of SARDINetV2 on the BigEarthNet-MM dataset which contains few samples of distorted images, the fine-tuning of the network on the Haiti dataset may have completely disturbed the process. Moreover, although the fine-tuning is performed on the entire network to allow for a deeper adaptation of its feature extraction capabilities in these distorted regions, the size of the dataset does not allow for accurate retraining. In addition, the translation network is not trained to meet the requirements of any downstream task and the resulting image may appear visually correct, but can lack relevance with respect to the actual land-cover and the downstream task. Indeed, the authors of [Wang et al., 2022a] have demonstrated that land-cover complexity can significantly affect the performance of the translation process. Geometrical distortions along with landslide collapse are typical land-cover complexities that can affect translation performance from a downstream task perspective.

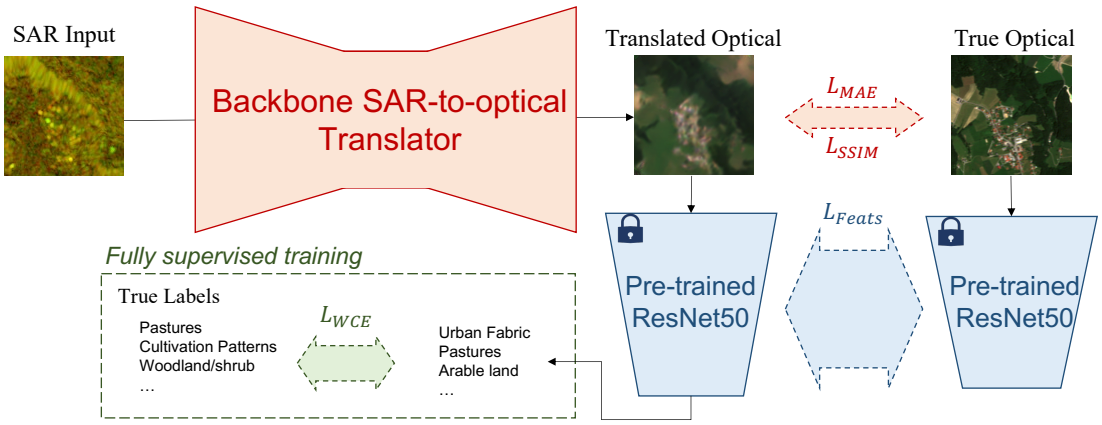


Figure 5.5: Training strategy based on the Translation-Classification Loss. The SAR input is first processed by a translator and compared to the true optical image. Both images are then evaluated by the frozen pre-trained classifier for feature comparisons. An additional fully supervised loss can be applied by comparing classifier results with the ground truth when land-cover labels are available.

5.4 Classification-Guided Modality Translation

Based on the counter performance observed in the previous section and its causes, this section proposes to explore the improvement of the relevance of the translations with respect to an auxiliary task. Indeed, since the diversity of the dataset is identified as a susceptible cause of the poor change detection performance, this means that SARDINetV2 does not correctly interpret the underlying land-cover. To avoid such a deficiency, this section proposes to force the network to reconstruct accurate patterns with respect to the actual underlying land-cover by leveraging the richness of BigEarthNet-MM [Sumbul et al., 2021]. To this end, a novel Translation-Classification Loss (TCL) is introduced and used to train different translator networks. These works proposed in [Bralet et al., 2024b] act as a prior for the training of a translator able to focus on the detection of landslides as detailed in the next Section 5.5.

5.4.1 Translation-Classification Loss

The use of an auxiliary classifier along with the main translator network has already been exploited to evaluate the translation performance of neural networks [Fang et al., 2023, Zhang et al., 2024]. Other works have even trained the classifier and the translator together to benefit from the complementarity of both tasks [Wang et al., 2022d, Zhang et al., 2020b]. Nevertheless, this combination is often applied to enrich the classification performance by considering the translation as an aside task [Ley et al., 2018]. Moreover, even when the translation is the core of the approach, some authors propose to classify the latent code of the generator to ensure reliable SAR feature extraction [Ley et al., 2018, Kandala et al., 2022, Wang et al., 2022d]. However, this strategy does not ensure accurate optical pattern reconstruction because SAR classification has been shown to be much less efficient than optical classification [Nava et al., 2021]. Therefore, the extracted features may not be relevant for the translation and the decoder responsible for pattern reconstruction does not benefit from the classification loss. Another peculiarity in the literature is that the auxiliary classifier is often trained together with the translator network. Such a strategy can lead to a lack of stability in the training since neither of the two networks is reliable, which means that evaluating the translation (resp. classification) task based on the results of the classifier (resp. translator) forces the network to rely on uncertain feature extraction.

For the above reasons, in this work, the training of the translator networks is based on an aside classifier network applied to the output of the translator. In addition, to avoid unreliable and unstable training, the classifier is pre-trained on real optical images to perform a multi-label land-cover classification. In this way,

the translator network is evaluated based on the relevance of the patterns reconstructed in the translated image. Furthermore, since the classifier is frozen to train the translator, the former evaluates the relevance of the reconstruction based on features that are expected to belong to real optical image distributions. Therefore, the land-cover classes in the reconstruction of the translator should not only be easily identifiable but also similar to the real optical images. Figure 5.5 shows the training routine of the translator as well as the different components of the proposed Translation-Classification Loss (TCL).

The goal of the TCL is to force the network to reconstruct images that are accurate with respect to the ground truth, but also to make best use of the auxiliary classification task. To this end, an input SAR image is fed into the translator network to produce a translated image $\mathcal{I}_T$. This image is then processed by the classifier to compute the prediction $\hat{y}$ of the land-cover reconstructed by the translator. In parallel, the true optical image $\mathcal{I}$ corresponding to the SAR image is also fed to the classifier for a deeper evaluation of the patterns reconstructed in the translated image $\mathcal{I}_T$. Based on this routine, the TCL measures both *translation errors* between $\mathcal{I}_T$ and $\mathcal{I}$, and *classification errors* between $\hat{y}$ and the ground truth y, comparing the behavior of the classifier depending on its input $\mathcal{I}_T$ or $\mathcal{I}$.

Translation errors: To evaluate the performance in translation, the loss terms are chosen to ensure both a radiometrically and structurally accurate optical-like image. From a radiometrical point of view, the authors of [Isola et al., 2017] devise that the Mean Absolute Error loss (L_{MAE}) performs better than the MSE loss. It not only reduces the blurry aspect of the final image but also limits the impact of outlier pixels on the training, *e.g.* the presence of small clouds. Then, the Structural Similarity Index Measure loss (L_{SSIM}) is also computed to ensure the transfer of the structural properties of the SAR images to the optical translated images. This loss has already demonstrated significant improvements as experimented in [Li et al., 2020]. Therefore, the combination of both loss terms aims to produce more relevant reconstructions with respect to the true image.

Classification errors: From a more global point of view, the classifier aims to detect whether the translated image contains features that are typical of the underlying land-cover or not. The prediction of the classifier $\hat{y}$ is compared to the ground truth labels y with the cross-entropy loss. However, since the training dataset may be unbalanced, each class is assigned a specific weight depending on the number of occurrences in the dataset. The resulting Weighted Cross-Entropy loss (L_{WCE}) is then more appropriate to make the network perform accurately on each class. In addition to this prediction loss and inspired by the work of [Wang et al., 2022a], another term is computed based on the features extracted by the classifier. Indeed, the authors of [Wang et al., 2022a] propose to evaluate the reconstruction of their translator by measuring the distance between the features of the decoder with the features of an aside optical autoencoder. Similarly, for the TCL, the classifier is used as an aside evaluator of the relevance of the reconstructed image. Therefore, the features extracted by the classifier from $\mathcal{I}_T$ are expected to be the same as those extracted from $\mathcal{I}$, the true optical image. Thus, a set of $N_{\mathcal{H}}$ hidden layers $\mathcal{H}$ is selected to compare their resulting features F_h with $h \in \mathcal{H}$ when the translated image and the true optical image are input to the classifier. This feature loss term (L_{Feats}), given in Equation (5.2), allows to evaluate the quality of the translation based on classification criteria at different resolution levels.

$$L_{Feats}(\mathcal{I}_T, \mathcal{I}) = \frac{1}{N_{\mathcal{H}}} \sum_{h \in \mathcal{H}} L_{MAE}(F_h(\mathcal{I}_T), F_h(\mathcal{I})) \tag{5.2}$$

This particular loss term has a second advantage over the WCE loss: it can be computed without ground truth labels. This peculiarity ensures that the TCL can be at least partially applied to any dataset containing SAR-optical pairs.

Translation-Classification Loss: The final TCL in Equation (5.3) is a linear combination of the four previous loss terms, considering λ_i where $i \in [0, 3]$ as weighted hyperparameters to balance the importance of each term for the optimization.

$$\begin{aligned} TCL(\mathcal{I}_T, \mathcal{I}, y) = \lambda_0 L_{MAE}(\mathcal{I}_T, \mathcal{I}) + \lambda_1 L_{SSIM}(\mathcal{I}_T, \mathcal{I}) \\ + \lambda_2 L_{Feats}(\mathcal{I}_T, \mathcal{I}) + \lambda_3 L_{WCE}(\hat{y}, y) \\ + L_{Spe} \end{aligned} \tag{5.3}$$

where the last term L_{Spe} refers to the specific loss term of a given architecture. More details on this term are given in the following Section 5.4.2.

Although the TCL can be applied to any classifier architecture, the ResNet50 architecture [He et al., 2016a] is chosen in this work because of its high performance for multi-label land-cover classification in [Sumbul et al., 2021]. Indeed, thanks to its residual connections, the network is able to extract accurate semantic features while avoiding information loss. The four post-downsampling layers are selected for the computation of L_{Feats}. Since the subsampling stages are applied at the end of each resolution block, they are intended to transfer only the most relevant information to the low-resolution layers. Therefore, these features are expected to be the most semantically informative and relevant for the final classification task, which justifies their use to evaluate the perspicacity of the translated images.

5.4.2 Backbone Translators

Since the TCL is applicable to any kind of translator architecture, its relevance has to be demonstrated on different models. Therefore, in addition to SARDINet and SARDINetV2 already described in Sections 4.3 and 4.5.1 and which do not require any specific loss, *i.e.* $L_{Spe} = 0$ in Equation (5.3), four additional architectures are experimented.

Based on the literature review given in Section 3.3, most translation architectures are based on an adversarial setting. Nevertheless, the previous chapter has already shown many issues with an adversarial training, especially the hallucinated artifacts. Therefore, to avoid such artifacts and ensure only the reconstruction of reliable patterns, all discriminators of the following architectures are disabled in our experiments. In Section 5.4.4, a brief investigation of the combination of the TCL and an adversarial training is given to provide insight into the relevance of the TCL for the generation of visually attractive images.

Based on the conclusions drawn in [Zhao et al., 2022a], and to be consistent with Chapter 4, Pix2pix [Isola et al., 2017] and CycleGAN [Zhu et al., 2017] are selected for the experiments. The Style-Recalibration Module (SRM)-based network in [Zhang et al., 2022] is the third architecture tested in this work to take advantage of its ability to distinguish the content from the style of the image. Finally, the Geometry-Consistent GAN (GCGAN) [Fu et al., 2019] is the last architecture used for the translation, as it is designed to be robust to image geometric transformations (*e.g.* rotation, translation, flip). Although the four architectures are already described in Section 3.3, for the sake of fluency, a short summary is given below:

- ***Pix2pix:*** Designed as an eight-stage U-Net [Ronneberger et al., 2015], the generator of Pix2pix is originally trained with the MAE loss. Since the latter is already part of the TCL, the term L_{Spe}^{P2P} is set to 0 in Equation (5.3).

- ***CycleGAN:*** This architecture is based on two different generators that translate a SAR image to an optical image and vice versa. The generators are designed as nine-block ResNets [He et al., 2016a] adapted for an unsupervised training. Therefore, two loss terms are used in [Zhu et al., 2017] to replace the supervision: the cycle-consistency loss L_{cycle} in Equation (3.22) and the identity loss L_{idt} in Equation (3.23). Therefore, the calculation of the specific loss of CycleGAN in Equation (5.3) is $L_{Spe}^{CG} = L_{cycle} + L_{idt}$.

- ***SRM:*** The SRM-based network has the same design as CycleGAN, except that each residual block contains an SRM to progressively transform the style of the input image to that of the expected output. Since it is also designed unsupervisely with two parallel generators, the specific loss term of the SRM-based network is the same as CycleGAN: $L_{Spe}^{SRM} = L_{cycle} + L_{idt}$.

- ***GCGAN:*** To ensure robustness to geometric transformations of the input image, the authors of GCGAN [Fu et al., 2019] propose to use an additional term to train a single generator network, also based on a nine-block ResNet architecture. Their geometry-consistent loss L_{geom} given in Equation (3.25) forces the network to be consistent from an image to its transformed version, *e.g.* rotation or flip. The final specific loss of GCGAN is then calculated as $L_{Spe}^{GCG} = L_{geom} + L_{idt}$

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	BigEarthNet-MM	Loss:	TCL	Metrics:	PSNR, MSE, FID, Accuracy
Input:	SAR (VH, VV)	Optimizer:	Adam		
Output:	Optical (RGB)	Batch Size:	16	SOTA:	Translators without TCL*
Size:	120×120 px♠	Learning Rate:	5.10^{-5}		
#Training:	34,412	#Epochs:	150		
#Evaluation:	7,520	$\lambda_0 : \lambda_1 : \lambda_2 : \lambda_3$ ♣:	$1{:}1{:}10^{-3}{:}10^{-3}$		
		Classifier:	ResNet50**		

Specificities

♠ Non-SARDINet architecture input 256×256px images, metrics computed after downsampling to 120×120px.

♣ The balancing weights are depending on the weights used for the original training.

* The unsupervised networks are trained supervisely for fairness by adding a MAE loss term to L_{Spe}

** The classifier is pre-trained under the configuration given in the text.

Table 5.3: Training configuration for the six networks trained with the TCL.

5.4.3 Translation Results

5.4.3.1 Experimental setup

The training of the six networks is conducted under the configuration mentioned in Table 5.3. Four peculiarities are to be mentioned during this training.

First, the size of the input images has been upsampled to 256×256 to be input to Pix2pix, CycleGAN, SRM and GCGAN due to the depth of these architectures. Indeed, based on the number of downsampling operations, the input size should be large enough to ensure the consistency of the latent space. Then, the metrics are computed by downsampling the translated images to the original resolution (120×120) to be comparable to the performance of SARDINet and SARDINetV2. Second, the hyperparameter weights λ_k used to balance the TCL are set differently depending on the original training of the models. However, the same ratio between the weights is computed, *i.e.* $\lambda_0 = \lambda_1 = \lambda_2.10^3 = \lambda_3.10^3$. The classification loss terms are given a lower weight to force the network to focus on reconstructing optical-like images before strengthening the relevance of the patterns by the mean of the classification task. The value of λ_0 is set to 100 for Pix2pix [Isola et al., 2017], 10 for CycleGAN [Zhu et al., 2017], SRM [Zhang et al., 2022] and GCGAN [Fu et al., 2019], and 1 for SARDINet [Bralet et al., 2022] and SARDINetV2 [Bralet et al., 2024b]. Third, in terms of state-of-the-art methods, the six networks are computed with their respective original losses. To ensure a fair comparison between supervised and unsupervised methods, the latter are supervisely trained by adding an MAE loss term. Therefore, each method is trained with the vanilla loss given in Equation (5.4).

$$L_{vanilla}(\mathcal{I}_T, \mathcal{I}) = L_{MAE}(\mathcal{I}_T, \mathcal{I}) + L_{Spe} \tag{5.4}$$

All networks are trained according to their Pytorch implementations available on Github[4,5,6]. Finally, the pre-training of the ResNet50 classifier [He et al., 2016a] is performed with the implementation available on Pytorch[7] on a larger dataset. Indeed, due to memory issues, the translators are only trained with a small subset of the BigEarthNet-MM dataset [Sumbul et al., 2021] while the ResNet50 is trained on $156,087$ samples and evaluated on $39,022$ images. The training is performed for 38 epochs using the WCE loss and an Adam optimizer with a learning rate of 1.10^{-4} and a batch size of 16. Such training allows the classifier to achieve more than 95% accuracy on the evaluation dataset, proving that the network succeeds in identifying the most relevant patterns for the multi-label land-cover classification task. It is important to mention that

[4]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix, last visited on July 17, 2024.

[5]https://github.com/summitgao/RSIT_SRM_ISD, last visited on July 17, 2024.

[6]https://github.com/hufu6371/GcGAN, last visited on July 17, 2024

[7]https://pytorch.org/vision/stable/models/generated/torchvision.models.resnet50.html?highlight=resnet50, last visted on July 17, 2024

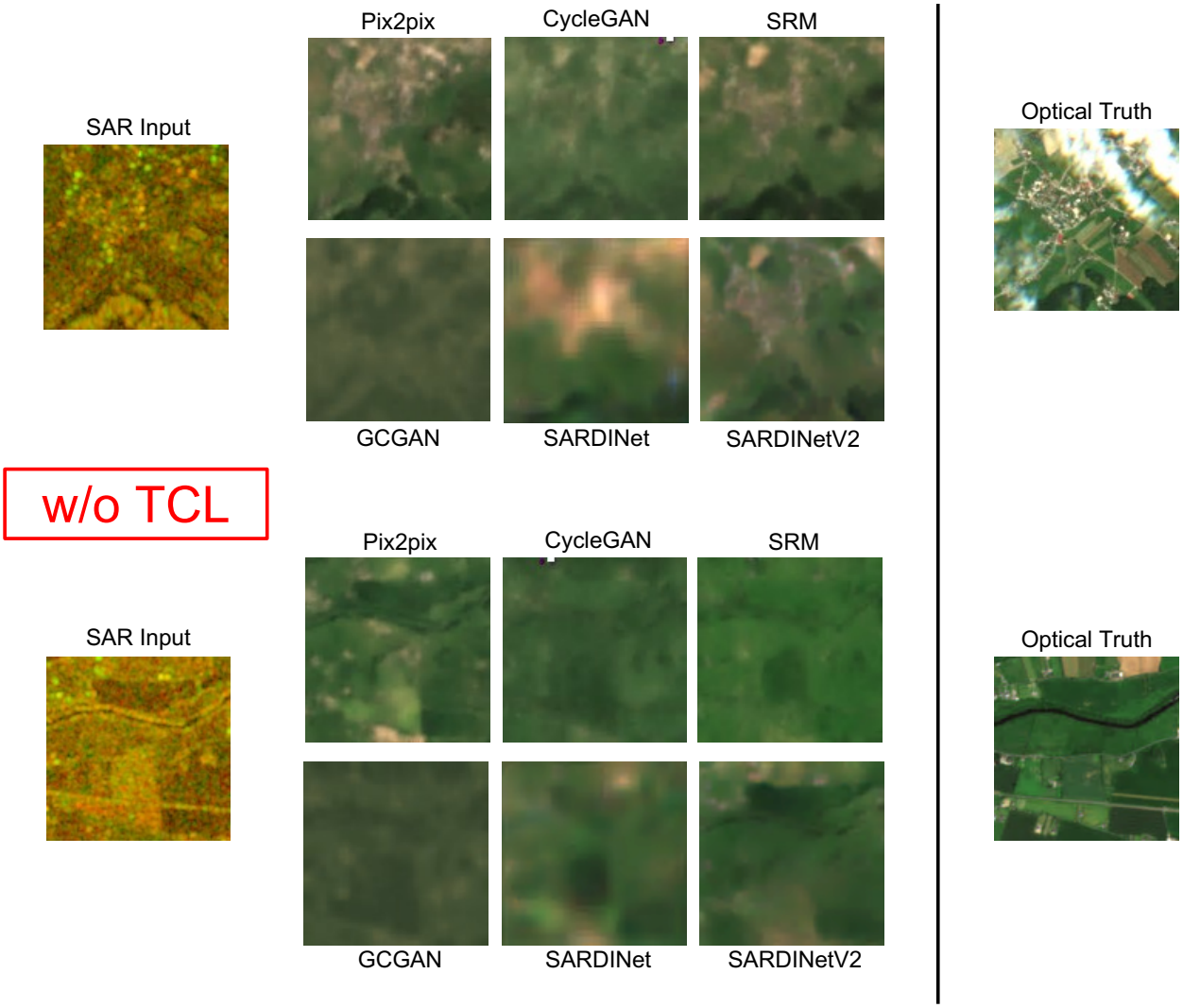


Figure 5.6: Qualitative performance of the generator-only networks when trained **without** the TCL. SAR images are colored with RG channels being VV and VH bands.

the use of the TCL has almost no impact on the training computational time. Indeed, with a NVIDIA RTX A6000 GPU, the training is only extended by 1 minute and 49 seconds per epoch.

5.4.3.2 Modality translation results

To better evaluate the relevance of the TCL applied to the different translator networks, Table 5.4 and Figures 5.6 and 5.7 illustrate their quantitative and qualitative performance. The first figure focuses on the performance obtained without the TCL while the second highlights the performance obtained with the TCL. In a few words, the most striking conclusion that can be drawn from these results is that the Translation-Classification Loss significantly improves the translation capabilities of neural networks.

The TCL has a strong impact on revealing the underlying structures and assigning more accurate patterns to the associated classes. Focusing on the top image of Figure 5.7, using the TCL to train CycleGAN helps to reveal the village and even to distinguish its different buildings. Similarly, on the second image and for almost all network architectures, the TCL helps to better reconstruct the presence of the small river which was poorly visible on the vanilla results of Pix2pix and SARDINetV2 in Figure 5.6. The TCL is also

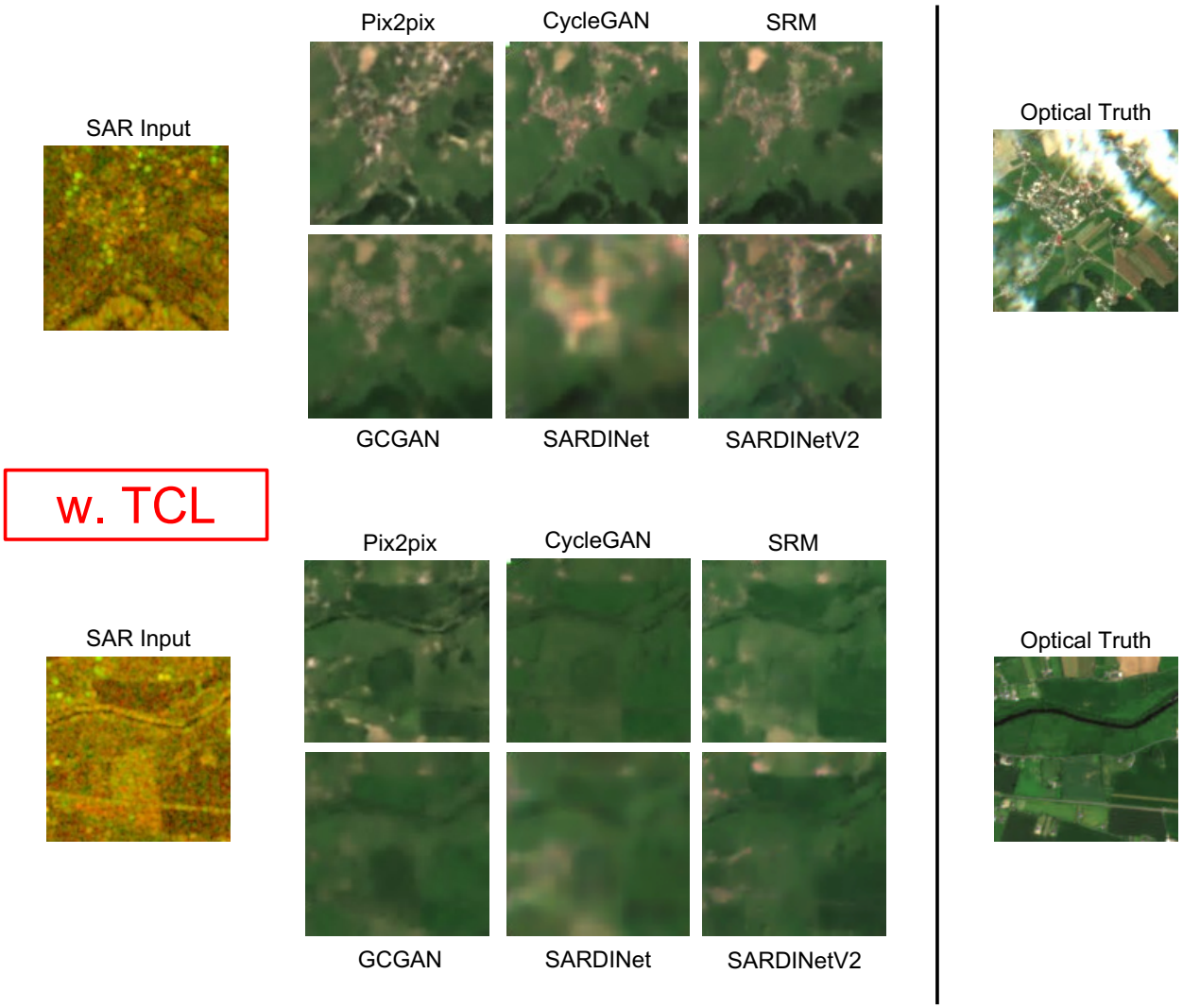


Figure 5.7: Qualitative performance improvement obtained **by using** the TCL for the training of generator-only translators. SAR images are colored with RG channels being VV and VH bands.

able to correct the errors of the vanilla networks visible in the first image where some forest patterns are initially reconstructed as buildings. These more accurate pattern reconstructions have a direct impact on the classifiability of the translations since the accuracy of Pix2pix and SARDINetV2 is doubled and that of CycleGAN and GCGAN is even multiplied by 8 and 4, respectively, in Table 5.4.

From an image quality point of view, two main characteristics should be taken into account: the general distribution of values of the image, *i.e.* its colors and brightness, and the contrast of the reconstructed image, *i.e.* the presence of details, textures and transitions between classes. On the one hand, the general ability of the TCL to rescale the distribution of the image is particularly visible in the reconstructions of GCGAN in Figure 5.7. Indeed, the vanilla reconstructions in Figure 5.6 are much darker with less color variations, making the fields and forests less distinguishable. However, by applying the TCL, there is a significant shift in color and brightness so that the village appears in the first image and a forest pattern can be obtained in the second image. This is also noticeable in the performance of the other networks: for instance, in the first image, the bottom forest patterns are much darker with the TCL allowing for better differentiation. Such visual evidence is also depicted by the quantitative metrics with a decrease of the MSE score by more than

	Without TCL				With TCL			
	PSNR ↑	MSE ↓	FID ↓	Acc ↑	PSNR ↑	MSE ↓	FID ↓	Acc ↑
Pix2Pix[Isola et al., 2017]	25.39	10.16e-3	4.34	15.47	25.55	**9.89e-3**	**3.81**	<u>34.68</u>
CycleGAN [Zhu et al., 2017]	25.17	13.26e-3	4.81	4.00	**25.67**	12.72e-3	<u>3.97</u>	31.82
GcGAN [Fu et al., 2019]	25.03	14.31e-3	6.66	6.94	<u>25.59</u>	12.19e-3	5.63	27.14
SRM [Zhang et al., 2022]	25.36	10.72e-3	4.36	11.54	<u>25.59</u>	<u>10.06e-3</u>	4.17	29.24
SARDINet [Bralet et al., 2022]	24.82	12.51e-3	4.64	17.22	25.25	11.85e-3	5.60	29.62
SARDINetV2 [Bralet et al., 2024b]	25.26	10.49e-3	5.07	13.34	25.32	10.32e-3	4.31	**35.88**

Table 5.4: Quantitative results comparing the impact of the TCL applied to standard translators. Best and second best results are displayed in bold and underlined, respectively. Accuracy is displayed in %.

2×10^{-3} for GCGAN and from 0.2×10^{-3} to 0.8×10^{-3} for the other networks. Similarly, the PSNR score increases for all methods indicating less noisy patterns and more accurate reconstructions. On the other hand, the use of the TCL allows a significant contrast enhancement without affecting the PSNR, contrary to the experiments in Section 4.5. The textures of the forest are more credible in the first image of Figure 5.7 where the TCL is used for the training and the transitions between field and forest or village and forest are also strengthened. From a general point of view, the overall images are less blurred than the reconstructions produced by the networks when trained with the vanilla losses in Figure 5.6. This implies the generation of more high-resolution details, such as the urban and small transitional forest patterns that appear in both images. This strong contrast enhancement is also reflected by the large decrease of the FID score in Table 5.4 which shows an average benefit of 0.7 points. Therefore, since the global and local patterns are improved thanks to the proposed TCL, it also explains the significant benefits obtained in terms of classification accuracy.

Interestingly, the TCL does not affect each network in the same way. Indeed, Pix2pix and SRM, which already reconstruct high resolution details, show particular gains in class transitions and color variations to better distinguish the different land-cover classes. The TCL even makes Pix2pix as the most beneficial since it achieves the best balance in terms of quantitative metrics in Table 5.4. Then, based on the performance of CycleGAN and GCGAN which lack contrast and consistency with their vanilla losses, the TCL allows the reconstruction of more detailed images and better balanced color distribution. Finally, SARDINetV2 is the network that shows the most stable quantitative translation metrics between the vanilla and the TCL trainings. The main impact of the TCL is a strong increase in terms of classifiability making SARDINetV2 as the best translator for the downstream classification task.

A single network shows a different trend both quantitatively and qualitatively. Indeed, the original SARDINet is still strongly affected by global blur despite the use of the TCL. The main reason for such a behavior is its inability to transfer high resolution features from the encoder to the decoder due to its architecture (see Figure 4.2). By losing these features in the successive MaxPooling stages, SARDINet decoder is not able to recover them. Conversely, thanks to the skip connections introduced in SARDINetV2, the latter can better exploit the structural properties of the SAR images, transfer them to the decoder and make them actually meaningful thanks to the TCL. Such a counter performance of SARDINet leads to two additional conclusions. First, the original architecture of SARDINet is better suited for high resolution images like those of SpaceNet 6 [Shermeyer et al., 2020] than for the medium resolution of Sentinel-1 and Sentinel-2 images belonging to BigEarthNet-MM [Sumbul et al., 2021]. Second, the TCL is not capable of inducing the recovery of high-resolution details in the reconstruction. The latter conclusion is crucial since it means that all patterns reconstructed by the other generators are not hallucinated but actually based on the properties of the input image.

Therefore, the TCL successfully brings more consistency, more relevance and more credibility to the SAR-to-optical translation without any artifact issue. The latter conclusion strengthens the interest of the TCL-based trainings over adversarial trainings.

Name	Translation		Classification		Metrics			
	L_{MAE}	L_{SSIM}	L_{Feats}	L_{WCE}	PSNR ↑	MSE ↓	FID ↓	Acc ↑
$\text{Config}_{\text{Init}}$	✓	✗	✗	✗	25.26	10.49e-3	5.07	13.34
x	✓	✗	✓	✗	25.21	10.38e-3	4.51	32.21
x	✓	✗	✗	✓	25.23	10.52e-3	4.94	27.82
x	✓	✗	✓	✓	25.25	10.73e-3	4.57	**39.96**
$\text{Config}_{\text{SSIM}}$	✓	✓	✗	✗	**25.46**	**10.06e-3**	4.37	15.80
$\text{Config}_{\text{Feat}}$	✓	✓	✓	✗	<u>25.43</u>	<u>10.30e-3</u>	**4.22**	28.29
x	✓	✓	✗	✓	25.41	10.47e-3	4.34	30.23
$\text{Config}_{\text{TCL}}$	✓	✓	✓	✓	25.32	10.32e-3	<u>4.31</u>	<u>35.88</u>

Table 5.5: Quantitative evaluation of the translations of SARDINetV2 trained under different TCL configurations. Best and second best results are displayed in bold and underlined, respectively. Accuracy is displayed in %.

5.4.3.3 Ablation of the TCL

To better understand the impact of the different terms of the TCL on the results, an ablation study is performed for each possible loss combination. Based on the aforementioned conclusions, SARDINetv2 gives the most stable results and is competitive with Pix2pix in terms of translation performance. Therefore, the latter is selected for the ablation study. For the sake of readability and to better highlight the most striking effects, only four configurations are considered for visualization purpose in Figure 5.8. Table 5.5 summarizes the quantitative performance of each possible configuration and denotes the components of the four selected configurations used in Figure 5.8.

- $Config_{Init}$: This configuration refers to the translation trained with the vanilla loss terms given in Equation (5.4).
- $Config_{SSIM}$: This configuration is based on the MAE and SSIM loss terms only. It helps to better identify the contribution of the SSIM in transferring the structural properties of the SAR images to the optical translations.
- $Config_{Feat}$: This configuration includes the feature comparison L_{Feats} in the loss along with the previous MAE and SSIM components. The objective is twofold. First to evaluate the impact of the feature comparison component on the structurality of the reconstructed patterns when combined with the SSIM loss term. Second, this configuration is the typical training routine that can be applied to an unannotated dataset. The performance of this configuration can be used to determine whether or not the TCL is applicable to any dataset containing SAR-optical pairs.
- $Config_{TCL}$: This final configuration contains all loss terms for a fully supervised training. The main goal is to study the impact of the final classification term on the performance of the translator.

The hyperparameters and overall training configurations are kept unchanged from Table 5.3 except for the balancing weights λ_i which are set to 0 for each symbol ✗ in Table 5.5, indicating the absence of the corresponding loss term.

By progressively introducing each loss term in the different configurations, their impact on the reconstructions can be better understood. In particular, the integration of the SSIM loss term L_{SSIM} in $\text{Config}_{\text{SSIM}}$ implies the most significant decrease in the FID score compared to the $\text{Config}_{\text{Init}}$ with 0.7 points. In addition, it achieves the best PSNR and MSE scores, illustrating that constraining the structurality of the reconstructed patterns is beneficial for most of the metrics. The same conclusion can be drawn based on the qualitative performance in Figure 5.8 since the buildings of the city are much better reconstructed, the transitions between the forests and the fields are strengthened and the thin transitional patterns are more

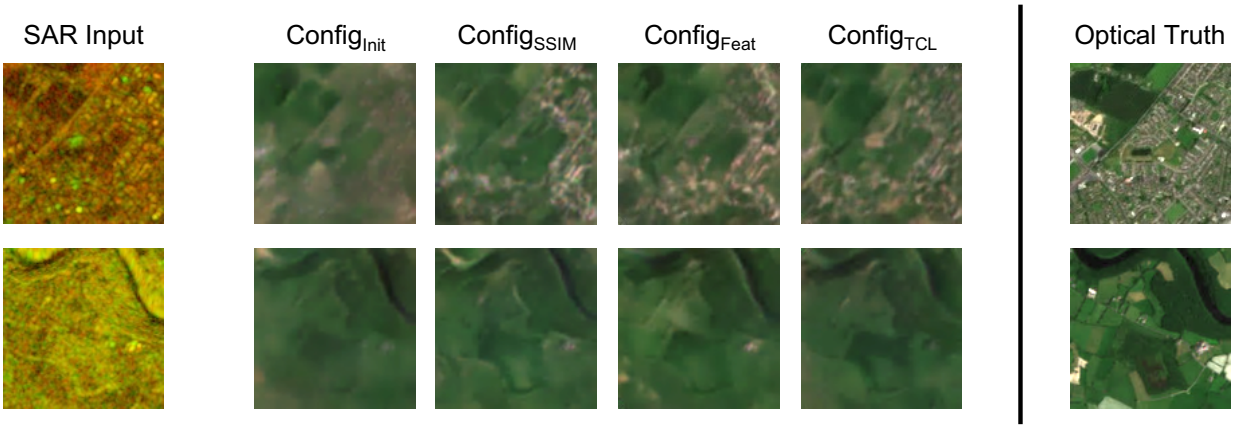


Figure 5.8: Qualitative comparison of the impact of the TCL components on the reconstructions of SARDINetV2. SAR images are colored with RG channels being VV and VH bands.

consistent with both the input image and the ground truth. However, based on the classification accuracy in Table 5.5, the relevance of the patterns is poorly increased by the L_{SSIM} term.

Here is the actual strength of the TCL. Indeed, by adding the L_{Feats} term in the Config$_{\text{Feat}}$, not only are the PSNR and MSE scores almost preserved, showing the second best performance, but the FID and accuracy scores are significantly improved with a decrease of 0.15 points and an increase of almost 13 points, respectively. The decrease of the FID score illustrates that this loss configuration strengthens the credibility of the images, as visible in Figure 5.8 where the reconstruction is less blurred, the city is enriched with more accurate buildings and the transitions from one class to another are even more enhanced than in Config$_{\text{SSIM}}$. On the other hand, the significant increase in classification accuracy indicates that these patterns are both more credible and more relevant. Config$_{\text{Feats}}$ shows the best balance of metrics over all tested configurations, which illustrates that despite the lack of labeled data in the training dataset, the TCL can still be applied and achieve actually interesting performance.

As a final evaluation, the integration of the fully supervised loss term L_{WCE} in Config$_{\text{TCL}}$ allows a significant increase in classification accuracy with an increase of more than 7 points compared to Config$_{\text{Feats}}$. Although Config$_{\text{TCL}}$ lowers the translation metrics, it still has the second best FID, third best MSE and fourth best PSNR scores. The significant increase in pattern relevance with respect to the underlying classes compared to the small losses in metrics remains an acceptable trade-off for better understanding the regions of interest. These lower translation metrics are reflected in the qualitative performance of Config$_{\text{TCL}}$ in Figure 5.8, as the city becomes blurrier and the small building patterns on the right of the second image are converted to forest patterns. Conversely, the discrimination of the forests is strengthened in both images, which may be caused by the imbalance of the labels in the training. Indeed, even if this phenomenon is limited by the weighting of the cross-entropy loss function, the classification task is still a global evaluation of the images, which favors large and recurrent patterns over scarce ones. A local evaluation, such as a land-cover segmentation task, may be an idea to solve such a problem. Nevertheless, the TCL shows significantly more interesting performance than all other configurations with respect to a downstream task. It not only allows the reconstruction of structurally consistent patterns thanks to the the L_{SSIM} and L_{Feats} terms, but also ensures that these patterns are coherent and relevant with respect to the underlying land-cover classes through the L_{Feats} and L_{WCE} terms.

5.4.4 TCL Outcomes

Based on the experiments conducted in this section, the counter performance observed in Section 5.3 can be better understood. Despite the accurate results of the vanilla training of SARDINetV2, the patterns do not match the peculiarities of the underlying land-cover which certainly limits their use for a downstream

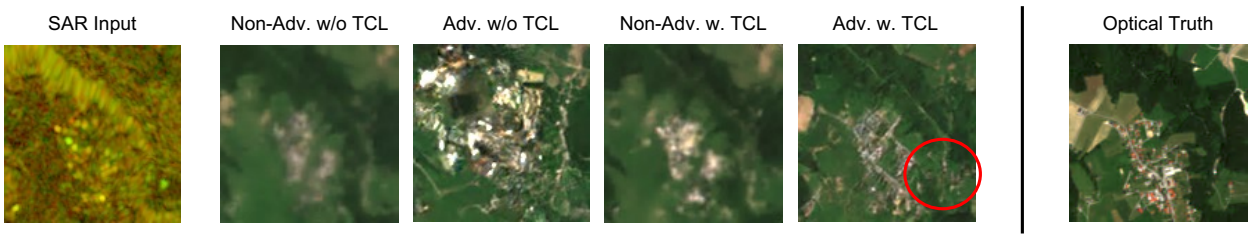


Figure 5.9: Comparison of the results obtained by Pix2pix [Isola et al., 2017] when trained adversarially and in combination with the TCL.

task. Moreover, the counter performance of the vanilla translator training is not due to the architecture of SARDINetV2 since all implemented translators are affected by such phenomena. However, by integrating the downstream task into the training of the translator network and by using a Translation-Classification Loss (TCL), the translation of each network is shown to be more relevant for land-cover classification. The translation metrics themselves are improved not only in terms of reliability but also in terms of credibility. A short experiment was computed to explore the regularization capabilities of the TCL by training Pix2pix [Isola et al., 2017] with both the TCL and an adversarial training. Figure 5.9 illustrates the performance obtained with such a configuration. Comparing the translations of the adversarially trained Pix2pix with and without the TCL shows that the latter significantly helps to reconstruct more reliable patterns. Nevertheless, apart from the strong regularization capability of the TCL, the hallucination drawbacks of the adversarial training are still visible in the reconstructed image, especially in the red circled area. Such consideration definitely lowers the reliability of the reconstructions compared to the non-adversarial training. Thus, the TCL is demonstrated to be advantageous in several aspects: credibility and relevance of the patterns, regularization of the network, inability to hallucinate, and land-cover reliability. These peculiarities of the TCL make it particularly interesting as a starting point for further fine-tuning on a different but related task, such as segmentation or change detection.

5.5 Change Detection based Modality Translation

Based on the high modality translation performance of the TCL, this section explores its extension to the detection of slope instabilities by replacing the classifier with the change detection networks explored in Section 5.3. Therefore, the TCL is converted to the Translation-Change Detection Loss (TCDL) following a similar procedure to fine-tune SARDINetV2. The change detection networks are then retrained with these new translated images to identify slope instability collapses. The TCDL is still at the experimental stage of the study. Therefore, this section relates the overall idea, modifications and peculiarities compared to the TCL before presenting the very first experiments in terms of modality translation and change detection.

5.5.1 Translation-Change Detection Loss

Inspired by the TCL routine proposed in Section 5.4, an extension to change detection is proposed here. Instead of computing a multi-label land-cover classification task, the ResNet50 is substituted by the optical-optical change detection network CDNet_S trained in Section 5.3. The main reason for choosing this Siamese architecture, despite the fact that it is not the most powerful network, is related to the feature loss term L_{Feats} needed for the computation of the Translation-Change Detection Loss (TCDL) described below.

Similar to the TCL, the TCDL consists of four different terms:

- L_{MAE} to evaluate the performance of the translation with respect to the post-event ground truth image,
- L_{SSIM} to evaluate the structural consistency of the translation with respect to the ground truth image,

- L_{WCE} computed on the change detection map or more specifically on the landslide change detection map resulting from CDNet_S,
- L_{Feats} which aims to compare the features extracted by the encoder of CDNet_S when inputting the translated image or the true post-event image.

Therefore, to focus only on the extraction related to the post-event image, the early fusion strategy cannot be used because the features of both the pre- and post-event images are directly merged. Conversely, the late fusion strategy trains a pre-event and a post-event encoder separately, which means that each encoder does not benefit from the training of the other. With the shared weight strategy, the Siamese configuration ensures that the extracted features actually belong to the input image, while using both images during the training.

Additional considerations should be taken into account for this particular training. Indeed, due to the complexity of the dataset described in Section 5.2, only a limited number of samples meets the expectations of the proposed TCDL-based training. First, the latter requires a clean optical image before the event to be processed by the change detection network, and second, it also needs a clean optical image after the event to compare the translation with the true image. To avoid these particular limitations, the cloud masks are used for the calculation of the losses. In particular, both the L_{MAE} and L_{SSIM} losses are computed only on pixels that are not affected by the clouds in the true post-event optical image $\mathcal{I}_2$ as given by the mask $\mathcal{M}_2$. Similarly, the L_{Feats} term is masked by the same cloud information downsampled according to the resolution of the features to be compared. The same masking operation is performed for the L_{WCE} loss, but using the cloud mask of the pre-event optical image given in $\mathcal{M}_1$. Indeed, all landslide pixels of the annotation map L that overlap the cloud map $\mathcal{M}_1$ are set to 0. In this way, the network is trained only to minimize the loss computed from the translated image $\mathcal{I}_T$ and the landslide prediction map $\hat{L}$ on reliable pixels. Such a training configuration implies that all images in the dataset can be used for fine-tuning SARDINetV2. Therefore, the final TCDL can be written mathematically as in Equation (5.5).

$$\begin{aligned} TCDL(\mathcal{I}_T, \mathcal{I}_2, \mathcal{M}_2, \hat{L}, \mathcal{M}_1, L) = \quad & \lambda_0 L_{MAE}(\mathcal{I}_T, \mathcal{I}_2, \mathcal{M}_2) + \lambda_1 L_{SSIM}(\mathcal{I}_T, \mathcal{I}_2, \mathcal{M}_2) \\ & + \lambda_2 L_{Feats}(\mathcal{I}_T, \mathcal{I}_2, \mathcal{M}_2) + \lambda_3 L_{WCE}(\hat{L}, L, \mathcal{M}_1) \end{aligned} \tag{5.5}$$

where the λ_i are the balancing weights as in the TCL (see Equation (5.3)), the L_{MAE}, L_{SSIM} and L_{WCE} are pointwise masked as in the MAE formula between two images of size $H \times W$ in Equation (5.6), and the L_{Feats} uses a downsampling operation "Down" to mask each of the N features of interest of an ensemble of layers $\mathcal{H}$ according to their resolution as given in Equation (5.7).

$$L_{MAE}(A, B, M) = \frac{1}{H \times W} \sum_{i}^{H} \sum_{j}^{W} |A_{i,j} - B_{i,j}|.M_{i.j} \tag{5.6}$$

$$L_{Feats}(A, B, M) = \frac{1}{N} \sum_{F \in \mathcal{H}} L_{MAE}(F(A), F(B), \text{Down}(M)) \tag{5.7}$$

Based on this TCDL, the network can be trained to specifically meet the requirements of the change detection task applied to landslide collapse identification.

5.5.2 SARDINetV2 Fine-tuning with TCDL

To be consistent with the experiments conducted in Section 5.3, the same hyperparameterization is used for the fine-tuning of SARDINetV2 as described in Section 5.3.2. The only differences are:

- The TCDL given in Equation (5.5) is used instead of the original fine-tuning loss in Equation (5.1). The weighting parameters are set to $\lambda_0 = \lambda_1 = 1$ and $\lambda_2 = \lambda_3 = 5$ to force the network to focus on the change detection task. In contrast to the classification task proposed in the TCL, the encoder of the CDNet allows the extraction of more high-resolution features. It also focuses on the extraction of landslide-like patterns, which were poorly considered by the original SARDINetV2 training.

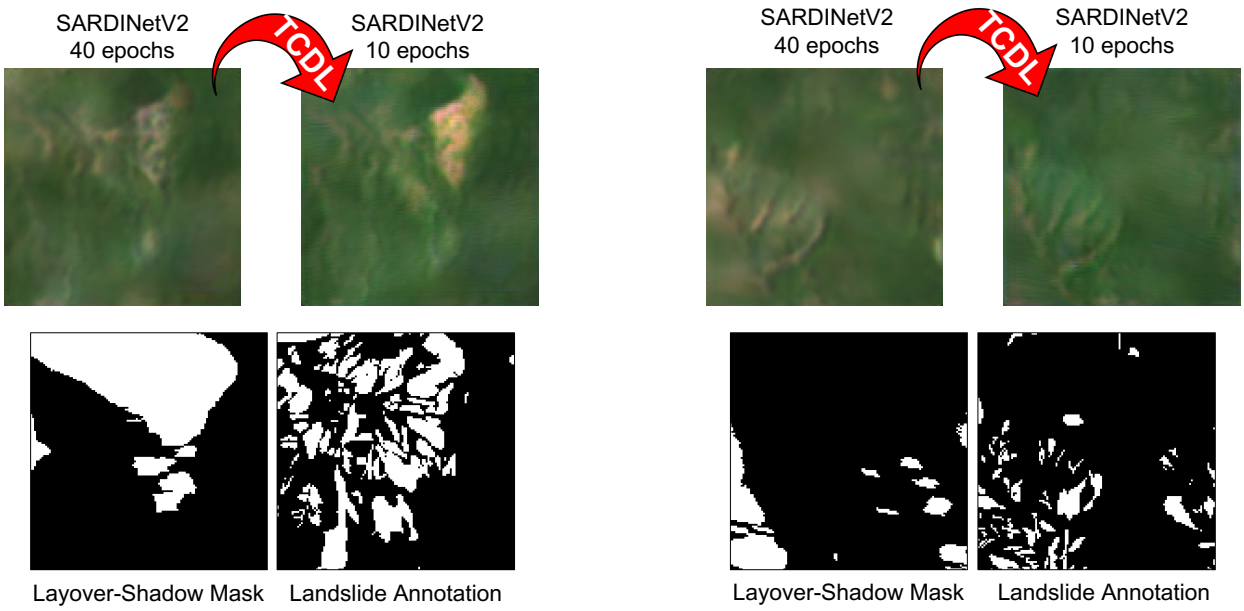


Figure 5.10: Qualitative comparison of the performance of the translators with and without the use of the TCDL for the fine-tuning stage.

- The fine-tuning is performed on 10 epochs only. Based on the evolution of the validation loss values during the training, the network is able to learn faster thanks to the use of the TCDL while keeping the same hyperparametrization. This shows that the TCDL helps to better focus on the relevant features. However, it also makes the network more prone to overfitting, which explains the lower number of epochs.
- To benefit from the TCL during the training, SARDINetV2 is initialized with the weights obtained in the previous Section 5.4. Therefore, the complexities of the land-cover are already integrated in the knowledge of the network. Such pre-training allows the fine-tuning of the network to better focus on domain adaptation and landslide pattern reconstruction.

The quantitative performance of the translation compared to the original training is given in Table 5.6. Figure 5.10 gives examples of the translated images resulting from SARDINetV2 fine-tuned with the TCDL compared to the vanilla training performed in Section 5.3.2. Several observations can be made regarding this comparison.

First, as highlighted in the translated image on the left, the landslide patterns are significantly enhanced when using the TCDL. Although the bottom landslides are still poorly identifiable, the large landslides at the top appear more contrasted which is encouraging for more accurate detections. This is even more supported by the quantitative metrics obtained in Table 5.6 where the FID score decreases significantly with the use of the TCDL compared to the original performance. In addition, the relevance metrics are maintained after the use of the TCDL which strengthens the interest of using the TCDL in this study.

Second, based on the right images, SARDINetV2 trained without the TCDL tends to generate several fuzzy patterns similar to landslides. However, these patterns are located in areas where the network lacks information due to the presence of geometrical distortions. Conversely, the use of the TCDL helps to better regularize such patterns in order to focus only on more reliable landslide patterns. Due to this behavior, we can hope a reduction of overestimation observed in Section 5.3.4 in the future change detection results. Nevertheless, due to the small size of the landslides in this image, they seem to be less visible in the translated images which may lead to a decrease in the precision of the change detection network.

	MSE ↓	SSIM ↑	PSNR ↑	FID ↓	MSE ↓	SSIM ↑	PSNR ↑	FID ↓
	Pre- and post-event images				Post-event images only			
w/o TCDL	11.80e-3	0.425	23.24	7.34	14.11e-3	0.388	20.33	8.71
w. TCDL	11.79e-3	0.416	23.25	5.97	14.24e-3	0.384	20.34	7.34

Table 5.6: Quantitative performance of the SAR-to-optical translation by comparing the use of the TCDL with the original fine-tuning.

5.5.3 Change Detection Results

The experimental setup used to compute the results in this section remains unchanged from that in Table 5.1. The unique difference is that the translations used to compose the optical-translation pairs come from SARDINetV2 fine-tuned with the TCDL given in the above Section 5.5.2. Similar to the original training, the use of CDNet_S quickly converged to a zero map, illustrating that the issues associated with the original training have not been corrected by the TCDL. The qualitative performance of the new TCDL-based CDNets is shown in Figure 5.11 for the same image as in Figure 5.4. The quantitative performance can be found in Table 5.7 where the two best monomodal CDNets and the two non-TCDL-based CDNets are recalled. Based on these preliminary results, partial conclusions can be drawn.

From a quantitative point of view, it is interesting to note that CDNet_L trained on the TCDL-based translations performs better than CDNet_E, while for all other configurations CDNet_E achieves the best results. This may be due to the increase in pattern credibility that makes the patterns more complex and more relevant for the change detection task. Therefore, extracting the features separately allows a better focus on the translated images than the early fusion which can lead to a dominance of the pre-event true optical image over the translated one.

Furthermore, compared to the training on the original optical-translation pairs, the TCDL-based CDNets show great quantitative improvements for almost all metrics, masked or not. In particular, the balance between the precision and the recall is much better which helps to increase the F1-Score of the results, reaching up to 0.36 for CDNet_L, increasing its interest compared to the use of monomodal SAR pairs. Nevertheless, the use of the TCDL-based translations still underperforms the results obtained with the monomodal SAR pairs.

Indeed, as depicted by the qualitative performance in Figure 5.11, the CDNets still largely overestimate the locations of landslides. Although the score map may look more focused on some particular landslides, the best F1-Score is achieved based on a very low threshold that causes such overestimations.

Network	mIoU ↑	F1-S ↑	Prec. ↑	Rec. ↑	Acc. ↑	Thresh.	#Epochs
Best Optical-Optical Change Detection							
CDNet_E	**0.384** / **0.276**	**0.545** / **0.420**	**55.11** / **55.10**	56.34 / 47.80	**94.73** / **84.74**	0.03 / 0.01	34
Best SAR-SAR Change Detection							
CDNet_E	0.258 / 0.251	0.404 / 0.396	34.08 / 34.56	51.32 / 47.80	87.18 / **84.74**	0.19 / 0.05	10
Optical-Translation Change Detection							
CDNet_E	0.200 / 0.202	0.330 / 0.332	25.76 / 24.19	54.05 / 57.93	77.35 / 75.40	0.07 / 0.06	19
CDNet_L	0.196 / 0.197	0.323 / 0.325	23.22 / 23.55	58.77 / 57.99	74.70 / 74.58	0.04 / 0.04	13
Optical-Translation Change Detection with TCDL							
CDNet_E	0.207 / 0.209	0.339 / 0.342	24.34 / 24.71	**60.52** / **59.86**	75.67 / 75.57	0.03 / 0.03	13
CDNet_L	0.223 / 0.223	0.361 / 0.361	27.15 / 27.45	57.14 / 56.26	79.09 / 78.92	0.04 / 0.04	13

Table 5.7: Quantitative performance of change detection networks trained on the translated images resulting from the fine-tuning of SARDINetV2 with the TCDL. Precision, recall and accuracy are given in %. Similar to Table 5.2, the performance are divided into "masked/actual performance".

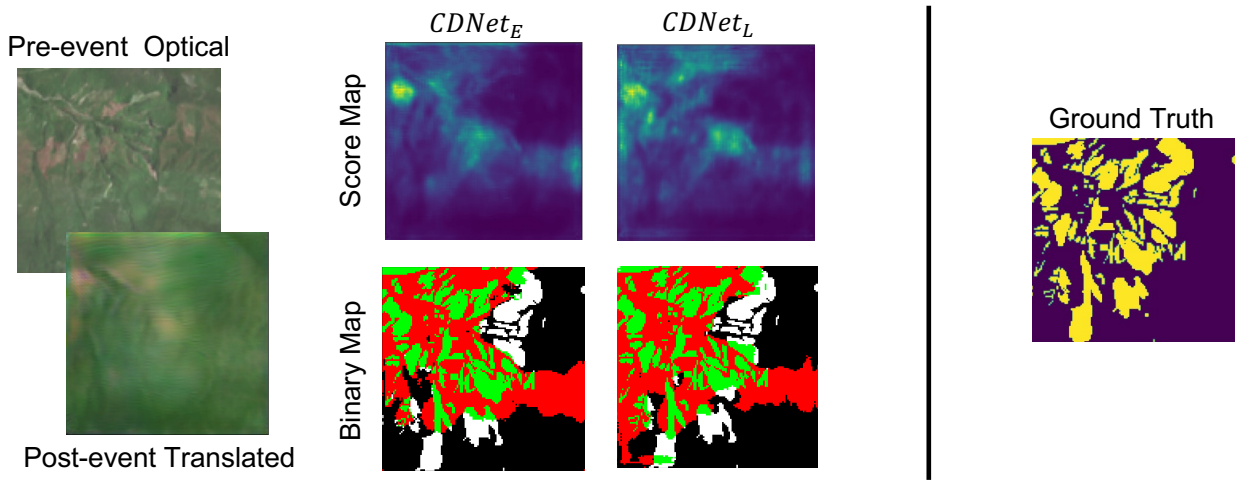


Figure 5.11: Qualitative performance of CDNet_E and CDNet_L trained with the translations of SARDINetV2 fine-tuned with the TCDL.

5.6 Conclusion

This chapter has mainly explored novel strategies to improve the relevance of SAR-to-optical translations obtained by SARDINetV2 with respect to a downstream change detection task. This change detection task is applied to a new dataset available on IEEE Dataport[8] and which associates bitemporal SAR and optical images with the delineations of landslides caused by the 2021 earthquake that struck Haiti. The dataset is first exploited to evaluate the relevance of translating post-event SAR images with SARDINetV2 for the change detection task. The use of straightforward change detection networks (CDNet) on pairs of pre-event optical image and post-event translated image has shown that the two images are too different from each other to be properly exploited by the CDNets. Several reasons for this counter performance have been identified including the dataset, the network architecture but also the quality of the translations with respect to the actual land-cover and the downstream change detection task. In this chapter, we proposed to focus mainly on improving the translation capabilities of the network, leaving the dataset and the change detection networks unchanged.

First, we proposed to address the improvement of the translations with respect to the underlying land-cover. To this end, we took advantage of the richness of the BigEarthNet-MM dataset [Sumbul et al., 2021] to learn the most characteristic land-cover features, which are intended to be further fine-tuned on the Haiti dataset and on the change detection task. BigEarthNet-MM images are provided with land-cover labels, which we propose to exploit by implementing a Translation-Classification Loss (TCL). This TCL involves the use of an auxiliary multi-label classifier network responsible for evaluating the performance of SARDINetV2 based on the classification ability of its translations and the similarity of the classification-related features to those extracted from true optical images. The TCL not only demonstrates more accurate and reliable translations for SARDINetV2 but also improves the performance of four other generators from the literature and helps to strongly regularize adversarial networks.

Based on this encouraging performance, the TCL is extended to the change detection task with the implementation of the Translation-Change Detection Loss (TCDL) which replaces the multi-label classifier with a Siamese CDNet. Therefore, by fine-tuning SARDINetV2 with the TCDL, the translator is expected to learn to generate more landslide-like patterns and be more relevant in the change detection task. Although the experiments are still in their early stages, the first observed performance with the TCDL is controversial. On the one hand, it shows a clear improvement in the credibility of the reconstructed patterns during the

[8]https://ieee-dataport.org/documents/multimodal-remote-sensing-dataset-landslide-change-detection-haiti, last visited on July 16, 2024.

translation process while maintaining their relevance. On the other hand, "maintaining" the relevance of these patterns is not sufficient to ensure accurate downstream change detection. Deeper optimization, more experiments and further comparisons with the literature [Luppino et al., 2021, Li et al., 2021b, Zhao and Wang, 2023] should be conducted to draw final conclusions regarding the relevance of the translations in the change detection process, the architecture of the network, and the complexity of the dataset due to its small size and challenging location.

Regarding the dataset, it is worth mentioning that the use of Sentinel-1 and Sentinel-2 images is very interesting due to their open accessibility but also due to their high revisit frequency. However, with the medium resolution of these images, the translation capabilities of the network are much more difficult, especially in mountainous regions where a large number of pixels are located in geometrically distorted regions. A promising approach would be to use higher resolution SAR images (*e.g.* TerraSAR-X or RADARSAT-2) to provide better insight into the underlying land-cover properties and to ensure even more reliable translations. The Sentinel-1 images would then serve as temporal information while the spatial information is provided by the other SAR image.

Regarding the translation process, we believe that a translation-based method can be applied to landslide change detection. However, the complete SAR-to-optical translation task may be extremely challenging and may not be necessary. An open research field from the literature is related to the *translation of the extracted features*. Two approaches that sound particularly promising are the use of hallucination networks, not to be confused with GANs, and the use of representation learning.

- **Hallucination networks** proposed in [Hoffman et al., 2016] are based on the use of modality specific feature extractors, *e.g.* a SAR and an optical one, as well as a hallucination feature extractor. Considering the case where an optical image is missing, the hallucination feature extractor is trained to mimic the feature extraction process of the optical-specific encoder but taking the SAR image as input. Therefore, by adding a decoder downstream of the feature extractors, change detection can be computed by comparing the features extracted from a pre-event optical image and a post-event SAR image. In contrast to the image-to-image SAR-to-optical modality translation process, this approach aims to translate the SAR features into optical features, which projects the ill-posed problem into the lower dimensional latent space.

- **Representation Learning** as explored in [Jain et al., 2021] consists of translating both image features into a common latent space. The so-called representation space is neither properly related to the optical nor to the SAR images but is a common representation of both images. In this way, the ill-posed problem caused by the translation can also be strongly mitigated, since the representation space is supposed to be suitable for both types of images.

Therefore, several research directions can be explored either by focusing on guiding the translations of the network, or by changing the point of view and focusing on feature translations instead of image translations.

Related communications

Journal

[Bralet *et al.*, CVIU 2024] *Translation-classification loss for SAR image understanding with deep learning*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E., submitted to Computer Vision and Image Understanding (CVIU), Elsevier - submitted on April 19, 2024

Dataset

[Bralet *et al.*, IEEE Dataport 2024] *Multimodal Remote Sensing Dataset for Landslide Change Detection in Haiti*, Bralet, A., Trouvé, E., Chanussot, J., Atto, A. M., on IEEE Dataport, July 13, 2024, doi: https://dx.doi.org/10.21227/4heb-7h07.

Chapter 6

Slow Moving Area Detection

6.1 Introduction

After the detection of post-collapse slope instabilities, this chapter focuses on their early detection, *i.e.* before they fail. Indeed, as stated in [Lacroix et al., 2020], evidence of slope instability collapse can be observed hours to months before the event occurs, which allows the event to be predicted and thus mitigated. Such clues are mainly visible in the form of small displacements, from millimeters to centimeters, often caused by triggering factors such as earthquakes or intense rainfall. These small displacements are closely related to the behavior of slow moving landslides and rock glaciers, which can move continuously for years without collapsing. Therefore, in this chapter, we propose to study the detection and localization of slow moving phenomena. The proposed methods can then be applied to detect precursors of sudden slope instability collapses. In Section 6.2, we propose the creation of a novel manually annotated dataset based on SAR Interferometry (InSAR) for the identification of centimeter-scale displacements. The use of the dataset with deep learning approaches for the segmentation task is described in Section 6.3. A more geoscience-friendly approach is proposed in Section 6.4 by introducing explainability during the training of a classifier network for slope instability detection and localization. Finally, in Section 6.5, the latter framework is used to extract additional segmentations and enrich the dataset introduced in Section 6.2.

6.2 Materials

The detection of slow moving instabilities cannot be properly investigated with the GRD Sentinel-1 and Sentinel-2 products. Indeed, the studied phenomena show movements ranging from millimeters to tens of centimeters per year while the Sentinel products have a 10-meter resolution in the ground geometry, which is too coarse for such fine displacements. This is the reason why, the focus of this study is set on exploiting interferograms generated from Sentinel-1 images. As mentioned in Section 2.4, these products are capable of revealing up to centimeter motions from SAR images. Nevertheless, despite the great efforts to build inventories of slow moving landslides and rock glaciers from interferograms [Bertone et al., 2022, Rouyet et al., 2021, Bodin et al., 2018, Bodin, 2020], no dataset has been built for deep learning approaches. Therefore, this section describes the different steps that we followed to create a new InSAR dataset for Slow SLIding area DEtection with machine learning (ISSLIDE). Figure 6.1 illustrates the three main steps of the process which are detailed below, namely the generation of the interferograms, the annotation process and the extraction of specific moving area patches.

6.2.1 Interferogram Generation

For the sake of reproducibility, low computational time and low external data requirements, the interferograms used in the dataset result from a standard yet effective pipeline derived from the "Command-Line

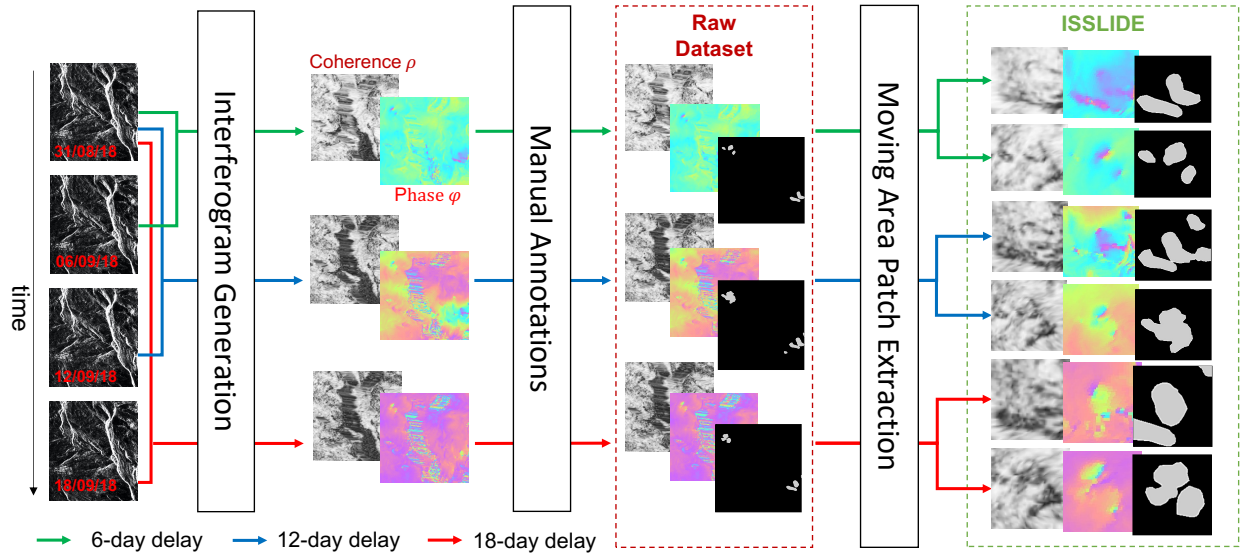


Figure 6.1: ISSLIDE generation pipeline. Each SAR image is coupled to another image acquired after 6, 12 or 18 days. The interferogram generation routine produces the coherence and phase difference maps which are then manually annotated according to expert guidelines. The resulting dataset is the raw ISSLIDE dataset which is further processed to extract interferogram patches surrounding the moving areas. The final Ready to Be Used dataset (R2BU) can then be used directly for deep learning training.

InSAR Tutorial" of the European Space Agency (ESA). The only requirement for computing the interferogram pipeline is the availability of a Digital Elevation Model (DEM). Since, the Shuttle Radar Topography Mission (SRTM) 1sec HGT product from [Farr et al., 2007] covers almost the entire Earth and is publicly available, it is chosen as the most relevant DEM in the following procedures. Contrary to the authors of [Thollard et al., 2021] and [Lazecký et al., 2020], the atmospheric disturbances are not corrected here. In addition, despite that Sentinel-1 products are released as dual-polarization images (VV and VH), only the VV polarization is used in this work. Contrary to the VH one, it has a strongest signal-to-noise ratio and ensures the generetation of more coherent interferograms. The proposed interferograms are computed with the Sentinel Application Platform (SNAP)[1] following eight processing stages.

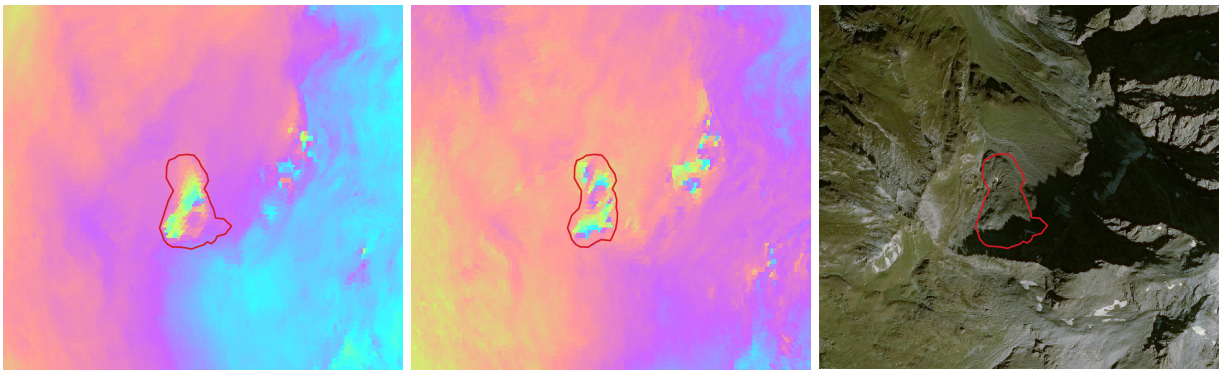
Image Calibration: This step reduces the radiometric bias in the SAR image and ensures that the pixel values actually correspond to the backscattering coefficient. Among the three proposed scalings, the σ_0 calibration is applied to the SAR images.

Application of the Orbit File: This step is used to ensure a finer estimation of the orbit of the satellite. Such file is directly provided by ESA.

Image Co-registration: This step is of major importance in the process because it allows the correct correspondence between the pixels of the two images. Indeed, a co-registration error will cause a shift in the two images resulting in an additional offset in the final phase difference map and lowering the coherence map. To ensure accurate co-registration, the second image is projected onto the first by the mean of a "back-geocoded" DEM, *i.e.* a DEM which is projected to the radar geometry. Obviously, the more accurate the DEM, the better the co-registration. Nevertheless, the 30 meter resolution SRTM 1sec HGT DEM [Farr et al., 2007] is used here for the reasons mentioned above. The DEM is bilinearly interpolated to match the resolution of the SAR images.

Interferogram Generation: After co-registration, the mandatory interferogram generation step described in Section 2.4.2 is computed. The neighborhood used to compute each pixel of the phase difference and coherence maps is set to a window of size 10×2 (range×azimuth) with a stride of one pixel so that the size of the interferogram is the same as its input. This window size is chosen to evaluate each component on

[1] https://step.esa.int/main/toolboxes/snap/, last visited on June 4, 2024.



(a) Bubble pattern on a 6-day interferogram

(b) Fringe pattern on an 18-day interferogram

(c) Optical image

Figure 6.2: Typical phase difference patterns identifying the presence of moving areas.

a nearly square area on the ground.

Topographic Correction: The topographic phase offset is a typical disturbance on the interferogram that can be evaluated based on a DEM. Therefore, using the same DEM as before, this offset is removed from the phase difference map.

Goldstein Filtering: The filtering step helps to reduce the overall noise of the phase difference map. By applying a standard and widely used two-dimensional Gaussian filter [Goldstein and Werner, 1998], the phase difference is cleaned up to make it easier to understand.

Multilooking: The multilooking step aims to subsample the image in a given neighborhood to produce images with a square pixel, *i.e.* with the same sampling in range and azimuth in ground geometry. The pixel values of the neighborhood are averaged to calculate the multilooked pixel value.

Orthorectification: To make the detection consistent with the optical images but also to allow the localization of the phenomena using geographic coordinates, the final interferograms are orthorectified. This step consists in projecting the image from the radar geometry to the ground geometry. Thus, by using the DEM, the over/under sampling issues due to geometrical distortions (see Section 2.3.2.2) can be compensated. The final projection coordinates refer to the World Geodetic System 1984 (WGS84) projection.

This series of operations allows the generation of the complex interferograms which can be handled as real and imaginary parts or as a phase difference map φ and a magnitude map ρ, commonly named as coherence map, as illustrated in Figure 6.1. Nevertheless, while unsupervised learning approaches can be trained using such maps, supervised methods require the dataset to be annotated in order to be effective.

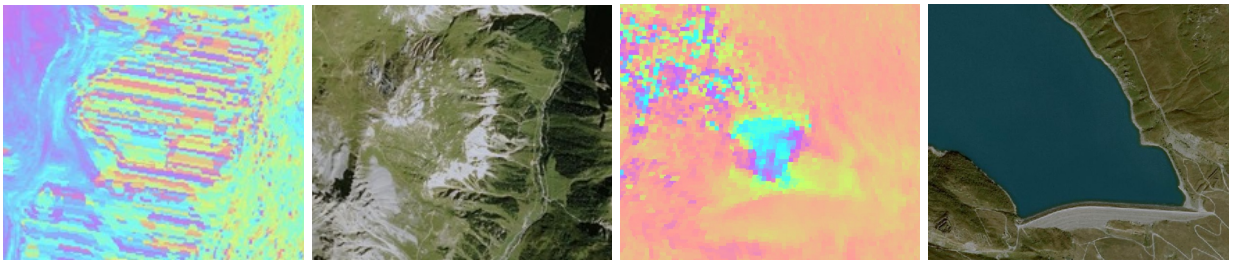
6.2.2 Annotation Process

The annotation strategy is designed to take the real characteristics of slow moving areas into account. An annotation procedure with guidelines for reliable detection is established in collaboration with geomorphological experts from the EDYTEM and ISTerre institutes. The manual annotation routine is performed with the QGIS software[2] and consists of two steps: the detection of suspicious patterns on phase difference maps and the validation of each detected area based on coherence, temporal and optical evidence.

6.2.2.1 Suspicious phase difference patterns

Two particular patterns are of interest for detecting slow moving areas in the computed interferograms: *bubbles* and *fringes*. The bubble patterns illustrated in Figure 6.2a are the typical phase patterns that identify the presence of a moving area. Indeed, from a geomorphological point of view, it reveals that the central area shows the maximum motion (in green in Figure 6.2a) and that it is surrounded by regions

[2] `https://www.qgis.org/fr/site/index.html`, visited on June 4, 2024.



(a) 6-day interferogram in a foreshortening region. (b) Foreshortening SPOT 6/7 image - 2018 (c) 6-day interferogram of a counterfactual pattern. (d) Counterfactual SPOT 6/7 image - 2018

Figure 6.3: Example of patterns easily confusing with moving area patterns.

which move slower (in yellow in the example). From an image processing perspective, the pattern breaks the homogeneity of the gradient of the difference map in the surrounding regions. On the other hand, the fringe patterns shown in Figure 6.2b highlight fast moving areas with a deformation which increases with the temporal baseline. Indeed, if the geomorphological interpretation is more challenging, these cyclic patterns reveal that the underlying displacements produce one or several phase rotations. Therefore, a raw estimate of the displacement can be computed by counting the number of fringes in the phenomenon. Typically, the motion in Figure 6.2b shows almost three fringes which corresponds to a displacement of $3 \times 2.7 = 8.1$cm in the direction of the Line Of Sight (LOS) between the acquisition of the two Sentinel-1 SAR images used in the interferogram.

An important characteristic in the interpretation of these patterns is their dependence on the temporal baseline of the interferogram. Indeed, both the bubble in Figure 6.2a and the fringe in Figure 6.2b were acquired over the same area during the same month. The only difference between these interferograms is their temporal baseline: 6 days and 18 days, respectively. This shows that annotating interferograms with multiple temporal baselines helps to better map the moving areas. Indeed, very slow displacements may be invisible on the 6-day interferograms but a bubble pattern can appear on the 18-day interferograms. The reciprocal is also to be considered: the fastest detectable moving areas can produce fringes in 6-day interferograms and be associated with noisy patterns in 18-day interferograms.

6.2.2.2 Validation process

A series of tests is applied to the suspicious patterns detected above. It consists in verifying the reliability of the detections from a signal processing and a geomorphological point of view.

Since interferograms are computed from SAR images, they are directly affected by the geometrical distortions. As depicted in Figure 6.3a, the patterns in the foreshortening areas could be assimilated as motion patterns. However, due to the poor information collected by the SAR images in such areas, they cannot be considered as reliable. Therefore, by computing the foreshortening and shadow maps during the orthorectification process as described in [Meier et al., 1993], the first test consists of masking all regions affected by layover and shadow distortions. The second test is based on the coherence map, which is used to validate the reliability of the phase difference map. Following the interferogram interpretation guidelines given in Section 2.4, the coherence of the detected region but also of its surroundings, is expected to be high. Indeed, although the area has moved between the two image acquisitions, the ratio between the displacement that occurred (centimeters) and the resolution of the SAR images (10 meters) is too small to cause a loss of coherence in the area.

From a geomorphological point of view, different characteristics of the phenomena of interest should be taken into account. Although both landslides and rock glaciers can have periods of inactivity, once a movement occurs, it should continue for several weeks in a row. Therefore, the third test expects the presence of the suspicious pattern in several interferograms with the same temporal baseline. Obviously, the more the pattern appears in the interferograms, the more reliable it is. Similarly, the presence of a motion pattern from one temporal baseline to another can be an indicator. Although such consistency is not

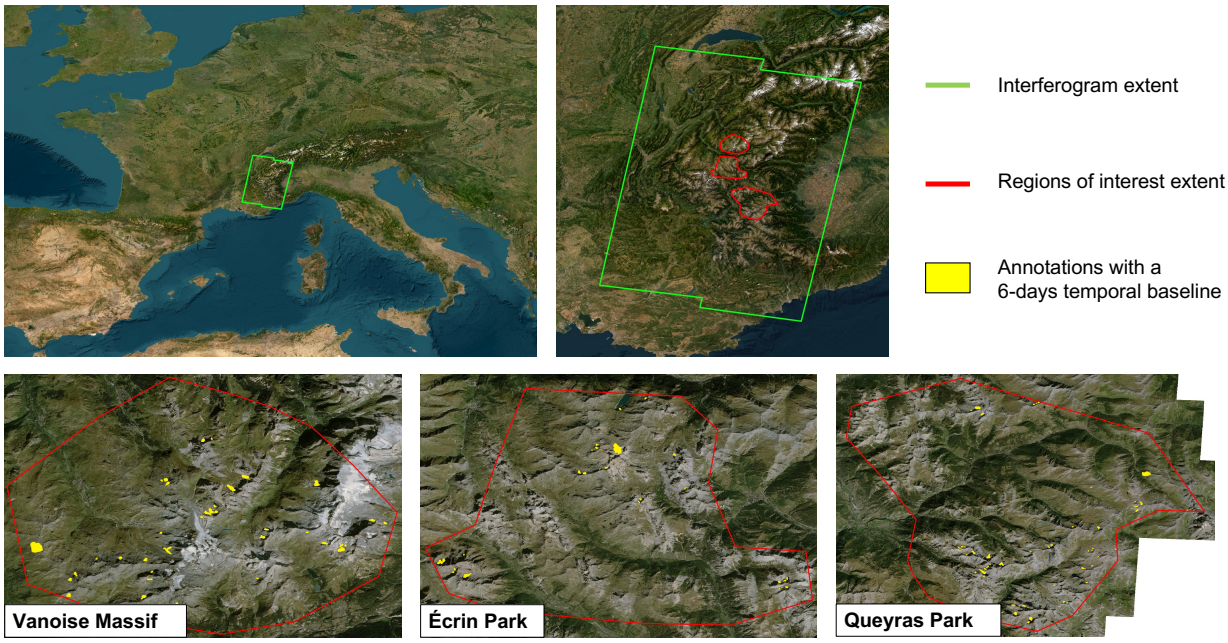


Figure 6.4: Region of interest considered for building the ISSLIDE dataset. A total of 136 moving areas are identified, and displayed in yellow, within 3 particular regions in the French Alps delineated in red. A total of 64 additional polygons not represented here are also annotated outside from these three regions.

mandatory as explained earlier, it is a clue that increases the confidence in the detection. The last test aims at the identification of geomorphological features on optical images. Indeed, it is first used to determine the type of detected phenomenon: the presence of a shear surface would identify a landslide, while ridges and furrows are characteristic of rock glaciers (see Section 2.2). The optical evaluation also allows the removal of outlier patterns like the one in Figure 6.3c. Despite the apparent phase difference bubble pattern, the latter is actually located in an artificial lake. Two possible causes for such a counterfactual are identified. First, the dam may suffer from internal deformations that are erroneously propagated to the lake due to the orthorectification step. Second, the water level may have risen or fallen between SAR acquisitions, and the double bounce of the wave against the water and the dam produces the phase offset similar to motion patterns. In any case, these types of patterns are not characteristic of slow moving slope instabilities and are therefore avoided during the annotation process.

6.2.3 ISSLIDE Dataset

The application of the interferogram generation pipeline and the annotation strategy developed in the previous sections results in a novel InSAR dataset for Slow SLIding area DEtection (ISSLIDE) detailed in [Bralet et al., 2024d] and available on IEEE Dataport[3].

The dataset is manually annotated over three specific zones of interest in the French Alps, namely the Vanoise Massif, the northeast of the Écrin Park, and the Queyras Park. Their location is delineated in red in Figure 6.4. To avoid the susceptible lack of coherence due to snowfall in the region in winter, the SAR images are acquired from July 2, 2018 to October 24, 2018. As Sentinel-1B was still operational in 2018, a total of 20 SAR images can be exploited with a delay of 6 days between each acquisition. The images are Level-1 Single Look Complex (SLC) products provided by ESA and acquired in Interferometric Wide (IW) swath mode only during the descending passes. Descending passes are privileged in the dataset because a

[3]https://ieee-dataport.org/documents/isslide-insar-dataset-slow-sliding-area-detection-machine-learning, last visited on June 28, 2024.

preliminary study of the ascending images revealed that they are not as rich as the descending images for this particular region. Of the three available swaths, only two are used to compute the interferograms as the third is acquired over non-mountainous regions. The green delineations in Figure 6.4 indicate the extent of these two swaths. Based on the annotation strategy guidelines, the interferogram generation pipeline is applied with three different temporal baselines, 6, 12 and 18 days, resulting in 19, 18 and 17 phase difference and coherence maps, respectively. The manual annotation is performed on each individual interferogram and ends with the detection of 44, 34 and 58 individual moving areas in the Vanoise Massif, Écrin Park and Queyras Park, respectively. The detected instability polygons are highlighted in yellow in Figure 6.4. These 136 annotations are complemented by 64 additional polygons outside these regions to enrich the dataset with a greater variety of shapes, phenomena and backgrounds, thus increasing the representativeness of the dataset.

To encourage enrichment of the dataset but also to make it easy to use in the context of deep learning research, two versions of the dataset are provided: a *raw dataset* and a *Ready to Be Used (R2BU) dataset*.

6.2.3.1 Raw dataset

The raw dataset is provided to encourage more detections based on the proposed interferograms. Indeed, the three annotated regions cover only 1910 km^2 of the 49,000 km^2 available on the complex interferograms. To open the exploitation of such a large unexplored area, the 54 interferograms, *i.e.* both the phase difference and the coherence map, are provided in the raw dataset together with the 54 shapefiles containing their respective individual annotated moving area polygons.

Each shapefile contains additional information about each of the 200 annotations to ensure the completeness of the dataset:

- A confidence index ranging from 1 to 10. Due to the subjectivity of the annotation process, the confidence index provides a measure of the uncertainty of the annotation. The latter is evaluated by the annotator himself depending on his confidence with respect to the criteria exposed in the annotation guidelines.
- The number of repetitions in the interferograms with the same temporal baseline. As stated in the guidelines, the repetitiveness between different baselines cannot reasonably be considered as mandatory evidence. Nevertheless, the repetitiveness of the patterns on interferograms with the same temporal baseline is actually an indicator of the reliability of the detection showing the continuity of the movement in time.
- A label identifying the category of movement observed. Three different classes are proposed based on the observed optical characteristics: *Landslides*, *Rock Glaciers* and *Unknown Phenomena*. The last class refers to the phenomena that show motion patterns on InSAR data but no clear geomorphological evidence on optical images.

These different attributes are informative about both the credibility and the uncertainty of each annotation. Such knowledge can be leveraged during the training of a deep network to weight the loss, avoid less confident annotations, perform weakly supervised training based on the confidence of the labels, or explore a curriculum learning strategy [Bengio et al., 2009, Liu and Zhang, 2022, Zhang and Ma, 2021]. The raw dataset is also made available to future users to enable custom patch extraction. Indeed, the R2BU dataset proposed below is a standard dataset that may not be compatible with very specific deep learning approaches. The raw dataset can therefore be exploited for such specific strategies.

6.2.3.2 Ready to Be Used dataset

The second dataset is built to make ISSLIDE easily applicable to deep learning approaches and to encourage the development of these methods in the context of moving area detection on interferograms. To this end, each interferogram is cropped around the annotated moving areas. The dataset aims to detect these phenomena regardless of their category, *i.e.* the annotation maps highlight the presence (1) or absence (0) of

movement in the region, without referring to whether the motion is due to a landslide or a rock glacier. To ensure the consistency of the dataset and to allow the use of multitemporal approaches, the patch extraction is identical for each of the interferograms.

First, all shapefiles of each interferogram for each temporal baseline are merged into a single global shapefile, which is then rasterized into a global annotation map. A distance transform is then applied to the global annotation map to localize the coordinates of each independent moving area by finding the local maxima of the distance map. Depending on the morphology and the size of the instabilities, multiple coordinates can be extracted from the same phenomenon. Yet, these duplicates are not disturbing for deep learning approaches as they act as a data augmentation strategy. Based on the extracted coordinates, a patch of 100×100 pixels is cropped from each of the interferograms and from the corresponding annotation map. Even though no motion is detected on a given interferogram for the given coordinate location, the patch is cropped and associated with a no-motion annotation map, *i.e.* an annotation map full of zeros. This routine allows the implementation of multitemporal strategies by avoiding to break the continuity of the annotation maps from one date to another. Similarly, a multi-baseline strategy can be implemented by stacking the interferograms with different temporal baselines to identify the evolution of the motion pattern on each of them.

Following this extraction pipeline, a total of 13,230 patches composed of the phase difference, coherence and annotation map are extracted. Among them, 8,701 patches show actual motion patterns. Thus, depending on the methodology to be implemented, one can consider using the whole R2BU dataset or filtering only the actual motion patches.

6.3 Standard Deep Segmentation Networks

Since the detection of moving areas on interferograms with deep learning approaches is poorly explored in the literature, this section proposes the application of standard deep learning segmentation networks to this task. It not only provides insight into the relevance and challenges of the ISSLIDE dataset with baseline performance [Bralet et al., 2024d], but also demonstrates the advantages and drawbacks of standard network architectures for the given task.

6.3.1 Standard Architectures

Five different networks are evaluated for the segmentation task. Two of them are standard models from the literature: Fully Convolutional Network (FCN) [Long et al., 2015] and DeepLabV3 [Chen et al., 2017]. Three other networks are implemented based on the standard U-Net architecture [Ronneberger et al., 2015]. The networks are described in detail below to better understand their strengths and weaknesses.

6.3.1.1 Fully Convolutionnal Network (FCN)

The FCN architecture introduced in [Long et al., 2015] is a direct segmentation extension of any standard classifier from the literature, *e.g.* AlexNet [Krizhevsky et al., 2012], VGG16 [Simonyan and Zisserman, 2015] or GoogleNet [Szegedy et al., 2015]. In this work, ResNet50 [He et al., 2016a] is selected as the reference classifier. The authors of FCN propose to remove all layers after the last convolutional layer. The features output from the latter are then input to an FCN head which corresponds to a series of Conv2D, BN and ReLU for feature refinement. The head ends with a Conv2D layer with a kernel size of 1×1, which can be associated with a linear combination of the feature maps resulting from the ReLU activation, to obtain a low-resolution segmentation map. The final segmentation map is obtained by bilinearly upsampling the output of the FCN head.

6.3.1.2 DeepLabV3

Similar to the FCN architecture, the DeepLabV3 network [Chen et al., 2017] exploits the feature extraction architecture of a standard classifier, here Resnet50 [He et al., 2016a], and applies a convolutional head to

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset:	ISSLIDE-R2BU	Loss:	BCE	Metrics♣:	HD, Dice, mIoU, F1-S, AUC
Input*:	ρ, $\cos(\varphi)$, $\sin(\varphi)$	Optimizer:	Adam		
Output:	Binary Segmentation	Batch Size:	8	SOTA:	FCN** [Long et al., 2015]
Size:	100 × 100 px	Learning Rate:	1.10^{-7} / 1.10^{-5}		DeepLabV3** [Chen et al., 2017]
#Training♦:	4,821	#Epochs:	600 / 300		UNet [Ronneberger et al., 2015]
#Evaluation♠:	572				ResUNet [He et al., 2016a]
					SepUNet [Chollet, 2017]
Specificities					
*, **, ♦, ♠, ♣, Blue/Green: For the sake of readability, the peculiarities are detailed in Section 6.3.2					

Table 6.1: Training configuration of the standard segmentation networks on ISSLIDE.

generate the segmentation maps. The peculiarity lies in the first stage of the DeepLabV3 head which consists of an atrous spatial pyramid pooling. The latter is a parallel feature extraction with different receptive fields to enhance the richness of the extracted information. The same input is passed to different Conv2D layers with a 3×3 kernel but with a different dilation ratio. The resulting features are then concatenated before being passed to an FCN-like head. Similarly, the final segmentation map results from the bilinear upsampling of the low-level decisions.

6.3.1.3 U-Net derivatives

The last three networks are inspired by the standard U-Net segmentation network [Ronneberger et al., 2015]. Each network is based on two encoding and two decoding stages with the same design: an alternation of Conv2D, BN and ReLU layers. MaxPooling and bilinear upsampling are responsible for the downsampling and upsampling stages, respectively. The two U-Net derivatives consist of the integration of residual blocks [He et al., 2016a] or separable convolutions [Chollet, 2017] in each encoder stage. Thus, while the standard architecture is called UNet, ResUNet and SepUNet refer to its residual and separable convolution variants.

6.3.2 Segmentation Results

6.3.2.1 Experimental setup

The five networks are trained according to the configuration detailed in Table 6.1. In particular, the interferograms are decomposed into three channels: the coherence (ρ), the cosine and the sine of the phase difference (φ). This separation limits the effect of the rotation of the phase on the network. In addition, as symbolized by ♦ in Table 6.1, to make the network even more robust, the training dataset is artificially augmented by adding a random phase offset to the phase difference map. This dataset consists of all annotated patches in the Vanoise Massif, the Écrin Park and outside the regions of interest. Moreover, only 80% of the interferograms are used for training in order to separate the training and evaluation datasets not only in space but also in time. Indeed, as materialized by the symbol ♠, the evaluation dataset is computed only on the patches of the remaining 20% of interferograms located in the Queyras Park. Finally, since the segmentation is applied to a single interferogram patch as input, only the patches showing actual motion patterns are selected for the training.

From an optimization point of view, since the FCN [Long et al., 2015] and the DeepLabV3 [Chen et al., 2017] are heavier than the three U-Net derivatives, their training hyperparameters are different to avoid overfitting. In particular, the Blue colored parameters are used for FCN and DeepLabV3 trainings while the Green parameters are used for the U-Net derivatives. In addition, since FCN and DeepLabV3 are widely used networks, they are trained both from scratch ($Model_S$) and with pre-trained initialization

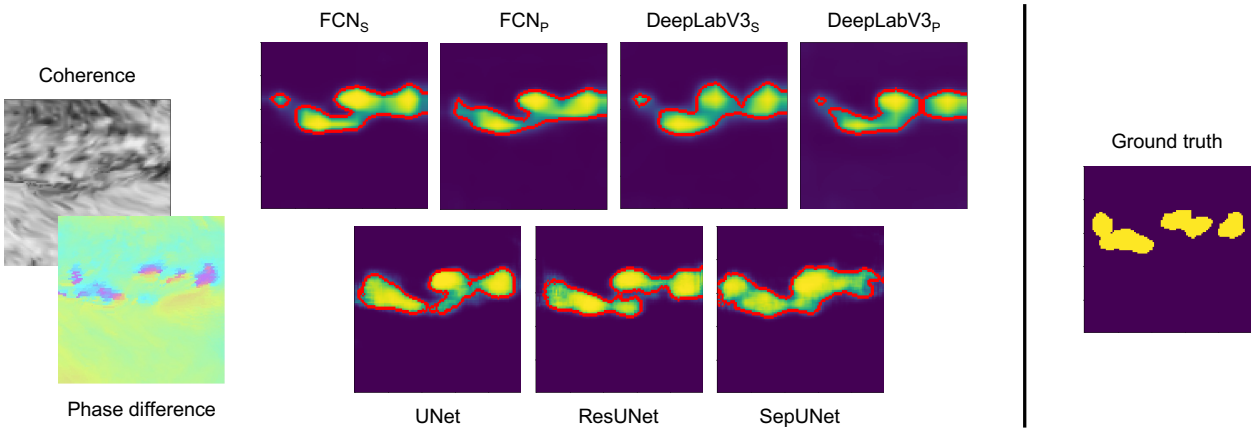


Figure 6.5: Qualitative results on ISSLIDE dataset for all the implemented standard segmentation models.

weights ($Model_P$) obtained with the PASCAL VOC dataset [Everingham et al., 2010] as illustrated by the ** in Table 6.1. The objective of this study is to evaluate the transferability of the networks from standard images to interferograms. The two models are trained under their Pytorch implementations[4] while U-Net derivatives are reimplemented from scratch. All networks are trained with a NVIDIA RTX A6000 GPU. The 100 epochs are performed in 4h for DeepLabV3, 2h30 for FCN, 2h for the UNet architecture and 1h30 for both ResUNet and SepUNet.

Finally, from an evaluation point of view, the Hausdorff Distance (HD), Dice index, Intersection of Union and F1-Score metrics require the binarization of the final probability map to be computed. Therefore, the symbol ♣ indicates that each individual network performance is computed based on a threshold given in the last column of Table 6.2 that maximizes the F1-Score.

6.3.2.2 Networks performance

The quantitative performance of the segmentation networks on ISSLIDE can be found in Table 6.2. From a qualitative point of view, due to the number of networks, two visualizations are proposed. Figure 6.5 presents the results on a single sample for all networks, while Figure 6.6 focuses on the best networks of each type on multiple samples to better evaluate their strengths and weaknesses.

The comparison of the two FCN trainings but also that of the two DeepLabV3 training figures out the transferability of the weights pre-trained on standard images to the InSAR data. From a quantitative point of view in Table 6.5, the pre-trained networks show a significant drop in performance compared to their raw counterparts. Indeed, from an ensemble point of view, the Dice and mIoU scores decrease by almost 0.02 points for the pre-trained FCN$_P$ and 0.01 for the pre-trained DeepLabV3$_P$. Similarly, the Hausdorff Distance increases by 1.3 and 0.5 points respectively, illustrating the under-segmentation produced by the pre-trained networks. Indeed, as visible in Figure 6.5, the pre-trained networks focus on the center of the moving areas with thinner segmentations than the networks trained from scratch. Such under-segmentation illustrates that the network cannot adequately leverage the knowledge transferred from the PASCAL VOC dataset and struggles to adapt them to the InSAR data. This behavior was expected since the three channels extracted from the interferogram do not represent the same information as an RGB image.

The comparison of the U-Net-like networks provides insights into the relevance of the different encoder blocks. In particular, based on the last rows of Table 6.2, ResUNet shows the most promising results as it achieves the second best overall performance. Although its quantitative metrics are close to those of UNet, the latter shows more uncertain segmentations as highlighted by the high threshold required to obtain the best F1-Score. By its side, SepUNet is significantly underperforming compared to its counterparts. Two main reasons can explain these results. First, separable convolutions imply a significant reduction

[4]https://pytorch.org/vision/main/models, last visited on June 6, 2024.

	F1-S ↑	AUC ↑	Dice ↑	mIoU ↑	HD ↓	Thresh.
FCN_S [Long et al., 2015]	**0.690**	**0.980**	**0.645**	**0.503**	**21.695**	0.27
FCN_P [Long et al., 2015]	0.677	0.976	0.628	0.486	22.951	0.31
$DeepLabV3_S$ [Chen et al., 2017]	0.673	0.978	0.630	0.486	<u>22.352</u>	0.36
$DeepLabV3_P$ [Chen et al., 2017]	0.666	0.976	0.620	0.480	22.887	0.32
UNet [Ronneberger et al., 2015]	0.675	0.977	0.636	0.497	25.912	0.44
ResUNet [He et al., 2016a]	<u>0.681</u>	<u>0.979</u>	<u>0.639</u>	<u>0.500</u>	25.921	0.35
SepUNet [Chollet, 2017]	0.651	0.972	0.606	0.466	28.026	0.32

Table 6.2: Quantitative comparison of the performance of the networks trained on ISSLIDE. Bold and underline results show best and second-best performance, respectively.

in the number of parameters. Since the implemented UNet architecture is already shallow, the use of separable convolutions makes it even less wide, limiting its ability to learn a wide variety of features. Second, the separable convolutions are originally implemented to better learn discriminative features from each independent channel. However, in these experiments, the three channels of the input patch are highly correlated. Indeed, both the cosine and sine channels are interdependent since they come from the same phase difference map, and the coherence map is an indicator of the reliability of the phase patterns. The independent feature extraction of these channels limits their interactions and thus leads to a decrease in performance.

From an architectural point of view, two comparisons can be made. First, comparing the performance of FCN and DeepLabV3 demonstrates that the former significantly outperforms the latter. Indeed, all classification, ensemble and distance metrics are improved with the FCN architecture. The main reason for these differences is the use of the atrous spatial pyramidal pooling layer in the DeepLabV3 architecture. Indeed, the resolution of the interferograms in this study is 10×10 meters. Therefore, atrous convolutions explore regions that may be far from the actual moving areas and thus extracting patterns that have no relation with the phenomenon. Indeed, as depicted in the first row of Figure 6.6, the centered motion pattern is not recognized by DeepLabV3 while it achieves coarse segmentations in the other two rows. The second architectural comparison consists in evaluating ResUNet with respect to FCN and DeepLabV3. It shows that the use of a decoder in ResUNet for the reconstruction of the segmentation map has significant advantages compared to the bilinear upsampling layers used in DeepLabV3 and FCN. Indeed, based on Figure 6.6, ResUNet allows a more detailed delineation of the boundaries (first row) as well as a better assessment of the uncertainty of the predictions from one pixel to another (second row). Although this uncertainty may decrease the ensemble metrics, it is more informative for future users than the coarse segmentations of its competitors.

However, the major drawback of these shallow architectures is their sensitivity to motion-like patterns with high coherence, such as the foreshortening area circled in orange in Figure 6.6. Indeed, ResUNet generates uncertain detections in these areas resulting in multiple small false alarms. This behavior explains the increase in Hausdorff Distance of 4.23 and 3.57 points compared to FCN and DeepLabV3, respectively, in Table 6.2. Yet, based on the pink circled area in the first row of Figure 6.6, all networks appropriately leverage the coherence map to avoid annotation of the counterfactual pattern. This has direct impact on the reduction of false positives in the segmentation map, which explains the high AUC values of all networks. This behavior seems to mimic the annotation guidelines listed in Section 6.2.2 which comforts on the ability of the networks to segment real moving areas.

6.3.3 InSAR Segmentation Outcomes

While the segmentation networks demonstrate convincing performance for the detection of moving areas on interferograms, a reasonable margin of improvement is still possible. Such a margin illustrates that

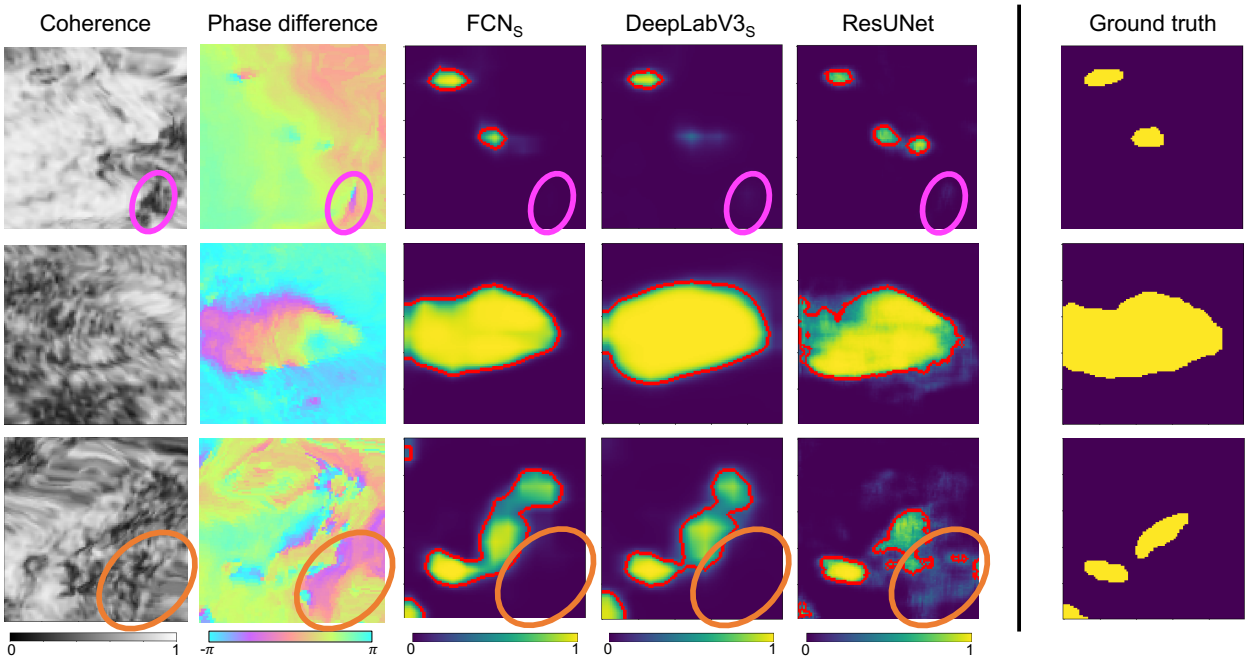


Figure 6.6: Qualitative results of the three best networks of each architecture type on different samples from the ISSLIDE dataset.

despite the wide applicability of these architectures on standard images, they are less efficient on InSAR data. Indeed, the peculiarities of interferograms make them even more difficult to interpret and understand not only for humans but also for deep feature extractors. But thanks to the availability of the ISSLIDE dataset, further studies can be conducted to adapt these standard architectures to this new type of data. In particular, the network should be better guided to distinguish the moving area patterns from counterfactuals and distorted regions despite the cyclic nature of the phase difference.

6.4 Localization through Learnt Explainable Detection

Guiding a neural network can take several forms in the literature: architectural modifications, multi-task learning or more complex loss computations. Sections 5.4 and 5.5 illustrate such guidance for modality translation by evaluating the translation based on a downstream multi-label classification and change detection task, respectively. Nevertheless, one of the biggest challenges with InSAR data is its understanding. Therefore, rather than guiding the network to obtain accurate results, we propose to train the network to explain itself in [Bralet et al., 2024a]. In particular, with the availability of the manual annotations in ISSLIDE, the core idea is to generate the localization map of moving areas based on the explanations provided by the Class Activation Map (CAM) of a straightforward classification network. This strategy has a number of advantages over standard segmentation approaches as given below.

6.4.1 Explainable Classification or Segmentation

The comparison between the use of an actual segmentation network and an explainable classifier is based on three main aspects: the algorithm, the output map and the user requirements.

From an algorithmic point of view, the classifier has great advantages over segmentation networks. As depicted in Table 6.3, the number of parameters of the classifier architectures are significantly lower than that of FCN [Long et al., 2015] and DeepLabV3 [Chen et al., 2017] with a ResNet50 backbone [He et al., 2016a].

	#Params (M)	FLOP (G)	Mem. Req. (MB)
Classifier Networks			
ResNet18 [He et al., 2016a]	11	1.81	44
ResNet50 [He et al., 2016a]	25	4.09	97
Segmentation Networks			
FCN [Long et al., 2015]	35	152.72	135
DeepLabV3 [Chen et al., 2017]	42	178.72	160

Table 6.3: Internal characteristics of classification and segmentation networks as implemented in PyTorch.

This characteristic is double-edged. On the one hand, the use of a larger number of parameters allows more variety in the feature extraction process, thus improving the performance of the network. On the other hand, using large networks on small datasets leads undoubtedly to an overfitting of the network on the training dataset as experienced in Section 6.3.2. In addition, such a number of parameters affects the number of Floating Point Operations (FLOP) of the networks. This metric measures the computational complexity of the network, which affects both memory and time requirements at inference time. In particular, FCN and DeepLabV3 are 38 and 44 times more expensive than ResNet50, respectively. This peculiarity severely limits their use in embedding systems that require both frugality and efficiency of the algorithms.

From a resulting product point of view, segmentation maps are *a priori* more interesting than standard classification approaches because they provide a pixel-wise decision. Nevertheless, the use of explainability to compute a localization map from the classification allows even richer results. First, the classifiers are able to detect the presence or absence of objects of interest with low computation requirements for fast decision making. Second, the use of the explainability can provide the most likely location of these detected instabilities to justify the decision. In addition, localization maps are more informative than segmentation maps because they can evaluate the uncertainty of detection at any given location while segmentation maps are mostly binary maps. In particular, slope instabilities are likely to cause their neighboring regions to move, creating uncertain boundaries that are easier to interpret with the localization map.

Finally, from a geoscientific point of view, segmentation networks have two major drawbacks. On the one hand, they are mostly considered as "black boxes" that make decisions without explaining the reasons for such decisions. Therefore, training a classifier to be explainable would ensure that non-experts in the field can better identify these reasons and more easily trust the decisions. In this work, the focus is on explaining the decision with the localization map, but the ECSPLAIN framework detailed below can be easily extended to other geomorphological attributes by considering more input data and/or other explanation strategies.

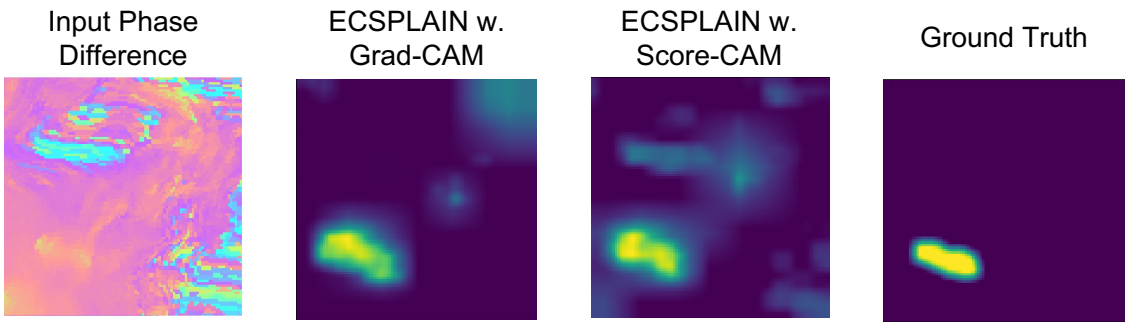


Figure 6.7: Localization maps generated by a classifier trained with the ECSPLAIN framework on a patch erroneously classified as "no motion".

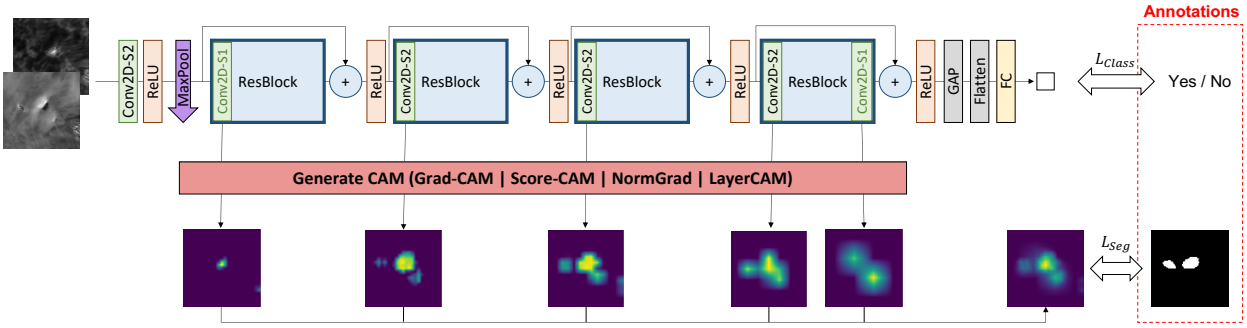


Figure 6.8: Overall scheme of the ECSPLAIN framework. A ResNet50 framework is used for classification of the real and imaginary parts of the interferogram patch. A multi-layer CAM is computed based on post-pooling convolutional features. The network is optimized based on both its classification and localization performance with the Binary Cross Entropy (BCE) loss.

On the other hand, the explainable classifications allow the future user to keep control over the final decision. Indeed, Figure 6.7 illustrates a specific case where the classifier missed the detection of the moving areas, *i.e.* the whole image is labeled as "no motion". However, despite the missed detection, the user can still compute the localization map to generate the regions most likely to contain a moving area. Conversely, "no motion" detection of a segmentation network results in a zero map. Therefore, the decision cannot be challenged by expert users while the explainable classifier can.

The three previous advantages of explainable classifiers over segmentation networks - *i.e.* the computational lightness, the localization map generation and the explainability property - encourage further improvements of localization performance to achieve competitive results with segmentation. In particular, constraining the explanation of the classifier to make it consistent with human expectations demonstrates promising performance but is still poorly explored in the literature.

6.4.2 ECSPLAIN Framework

The general idea of constrained explanation is quite simple: during the training phase of the network, each decision is explained and the loss calculates the distance between the computed explanation and the expected explanation. In this way, the network is trained to explain its decisions as a human would. But depending on the task, the explanation method and the input data, the implementation can be more challenging. This section details the framework based on an "Explainability Constrained-claSsifier for Pairing the detection and Localization of moving Areas on INterferograms" (ECSPLAIN) proposed in [Bralet et al., 2024a]. The overall design of the framework is depicted in Figure 6.8.

Since the method is applied to interferogram patches, the rotation of the phase difference map is still an issue. In Section 6.3, this is overcome by computing the sine and cosine of the phase difference map concatenated to the coherence map, resulting in three channels $[cos(\varphi), sin(\varphi), \rho]$. Similarly, the complex interferograms are fed into the network by separating the real ($\rho cos(\varphi)$) and imaginary ($\rho sin(\varphi)$) parts. This separation is motivated by two reasons: first, it slightly reduces the memory requirements of the network and second, it allows the coherence map to be directly integrated to the input data. Thus, the network does not need to interpret the coherence map as a measure of the reliability of the phase difference since it is directly included in the input. The focus of the network is then directly set on the extraction of motion patterns.

The ResNet50 [He et al., 2016a] architecture is selected as the classifier for this study. This choice is made because ResNet50 is widely used in computer vision for classification but also for segmentation tasks by acting as a backbone encoder. The Pytorch implementation of the network is slightly modified to take two-channel images as input and output a single scalar value for the detection of moving areas. In addition, the Batch Normalization layers are removed to ensure better stability of the training. Based on the conclusions of Section 6.3, using pre-training weights computed from a computer vision dataset to

initialize the networks degrades the final performance. Thus the networks are trained from scratch to be truly dedicated to interferogram inputs.

The explanation method responsible for the generation of localization maps is the Class Activation Map (CAM). Four of them are tested in the framework: Grad-CAM [Selvaraju et al., 2017], Layer-CAM [Jiang et al., 2021], NormGrad [Rebuffi et al., 2020] and Score-CAM [Wang et al., 2020a]. The mathematical formulation of these CAMs is given in Section 3.5.1. To ensure the reconstruction of more detailed localization maps [Liu and Zhang, 2023, Fu et al., 2023], the CAMs are generated on an ensemble of target layers T and combined into a multi-layer CAM as in Equation (6.1).

$$CAM = \frac{1}{\#T} \sum_{L \in T} Up(\frac{ReLU(CAM_L) - min_L}{max_L - min_L}) \tag{6.1}$$

where $Up()$ is a bilinear upsampling operation to fit the size of the input image and to ensure that all CAMs have the same resolution. Then, $ReLU()$ and min-max normalization are applied following the standard CAM generation algorithm guidelines. The former is applied to keep only the regions with a positive localization score, while the normalization avoids the dominance of the layers with the largest number of feature maps. For this study, $\#T = 5$ as the four post-pooling convolutional layers and the very last convolutional layer are selected for CAM computation as visible in Figure 6.8. The four layers are selected after the pooling operations because the latter ensures the extraction of the most relevant features from the previous layer, which is what the explanation strategy aims to understand.

Contrary to the standard use of explainability methods, the ECSPLAIN framework proposes to generate the multi-layer CAM during the training phase. In this way, the localization can be evaluated by the loss and the network can be trained to better explain its decisions and match end-user expectations. In particular, the ECSPLAIN framework constrains the computed CAM to match the segmentation of the actual moving area which is delineated in ISSLIDE (see Section 6.2). Thus, similar to standard segmentation losses, the Binary Cross Entropy (BCE) loss, defined in Equation (3.20), is used to evaluate the relevance of the localization map with respect to the annotations. Coupling this segmentation loss (L_{Seg}) with the standard classification loss (L_{Class}) - also the BCE loss - applied to the output of the ResNet50 classifier forces the network to maintain consistency between its detections and explanations. Equation (6.2) illustrates the computation of the loss from the set of predicted labels and localization maps $[\hat{y}, CAM]$ and the set of ground truth labels and annotation maps $[y, A]$.

$$L([\hat{y}, CAM], [y, A]) = L_{Class}(\hat{y}, y) + \lambda L_{Seg}(CAM, A) \tag{6.2}$$

where λ is a hyperparameter weight used to balance the contribution of the segmentation and classification tasks.

6.4.3 Slow Moving Area Localization Results

The ECSPLAIN framework raises several questions that should be addressed in this results section. Therefore, after detailing the experimental setup and its peculiarities in the following paragraphs, four main ideas are discussed based on the results:

- Is the localization task as powerful as the segmentation task?
- Is an explainability-constrained training actually more relevant than explaining *a posteriori*?
- Which CAM is most relevant for the localization of moving areas?
- Is a network trained with a given CAM explainable with another CAM?

All these issues are discussed separately in the following to get a complete overview of the strengths, weaknesses, applicability and limitations of the ECSPLAIN framework.

Training Configuration					
Dataset Configuration		Optimization Hyperparameters		Evaluation Setup	
Dataset*:	ISSLIDE	Loss:	Equation (6.2)	Classif. Metrics:	Accuracy, Precision, Recall,
Input:	$[\rho cos(\varphi), \rho sin(\varphi)]$	Loss weight λ:	1**		F1-Score
Output:	Label + Map	Optimizer:	Adam	Segm. Metrics†:	Dice, mIoU, HD, AUC, MAE
Size:	100×100 px	Batch size:	8		
#Training:	8,483	Learning Rate:	1.10^{-4}	SOTA♠:	FCN [Long et al., 2015]
#Evaluation:	1,638	#Epochs:	194		DeepLabV3 [Chen et al., 2017]
					ResUNet [He et al., 2016a]
					A posteriori CAMs
Specificities					
* The R2BU dataset is used for positive samples and additional crops from the raw dataset are used for negative samples. The same data split and augmentation strategies as in Section 6.3.2 are used.					
** λ is selected to 1 since it showed the best performance.					
♠ SOTA methods are trained and evaluated with a different hyperparameter setting. More details in the text.					
† The MAE is used to evaluate the uncertainty modeling capabilities of the networks. The remaining segmentation metrics use a threshold which maximizes the Dice score. Note that the computation of the Dice is different from the F1-Score since the it also rewards the prediction of maps full of zeros when they should be.					

Table 6.4: Training configuration of the ECSPLAIN framework.

6.4.3.1 Experimental setup

The overall configuration used for training the ResNet50 classifiers under the ECSPLAIN framework is given in Table 6.4. The only peculiarity not mentioned is the fact that the training of ECSPLAIN with Score-CAM is performed with a *feature dropout*. Indeed, due to memory limitations, instead of computing the score for each of the N_L feature maps, only 10% of them are randomly selected for each batch for CAM generation. However, at inference time, all feature maps are used to obtain the final Score-CAM.

Among other peculiarities, the R2BU dataset of ISSLIDE [Bralet et al., 2024d] described in Section 6.2 is enriched. Indeed, due to the detection task, the classifier training requires both positive and negative samples. The latter are randomly extracted from the interferograms provided in the raw dataset. From a data splitting and data augmentation points of view, the same strategies are used as in Section 6.3.2.

A novel hyperparameter λ balancing the importance of the segmentation loss is introduced in this experiment. The latter was experimented under different configuration ranging from 0.1 to 100. It appeared that giving the same importance to both tasks, *i.e.* $\lambda = 1$ results in the best performance.

The three best segmentation networks are also retrained on the dataset for baseline comparisons. Yet, due to their size, they cannot be trained with exactly the same hyperparameterization. In particular, the learning rate of FCN [Long et al., 2015] and DeepLabV3 [Chen and Sun, 2023] is set to 1.10^{-7} and that of ResUNet to 1.10^{-5}. The three networks are then trained for 1000 epochs while taking care not to overfit. Further experiments are also performed with *a posteriori* CAMs. They consist in training the ResNet50 classifier only for the detection task, *i.e.* $\lambda = 0$ in Equation (6.2), and generating the CAMs only at inference time.

Finally, from a metrics point of view, Dice, mean Intersection over Union (mIoU) and Hausdorff Distance (HD) are computed after binarizing the localization map with a threshold that maximizes the Dice score. Importantly, for this particular experiment only, the Dice score is computed differently from the F1-Score since the maps that are predicted as full of zeros are not rewarded in the computation of the original F1-Score. The Dice used in this work rewards such predictions when the ground truth is actually a zero-map. The computation of the Dice is obtained based on the implementation of the miseval library[5]. Conversely, the MAE score is computed with a blurred version of the annotation map to evaluate the uncertainty modeling ability of the different networks. Finally, the classification metrics of the segmentation-only networks are

[5]https://github.com/frankkramer-lab/miseval, last visited on July 13, 2024

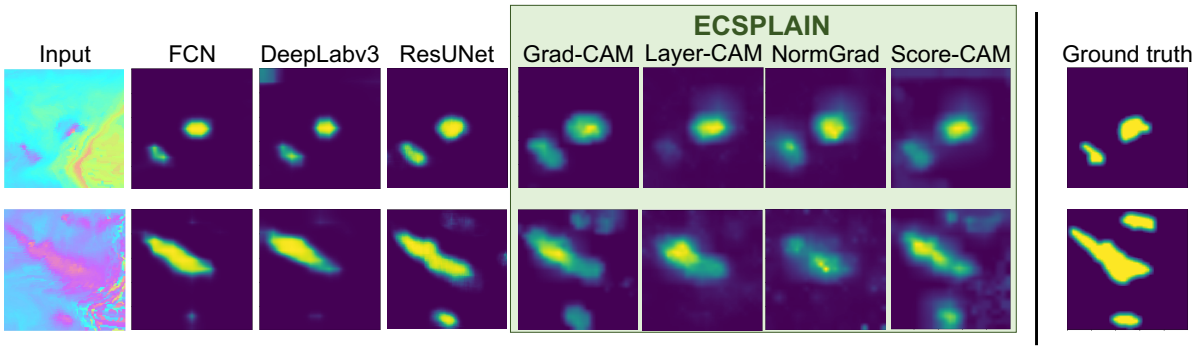


Figure 6.9: Qualitative comparison of the segmentation network and the ECSPLAIN framework trained with the different CAMs. For the sake of readability, only the phase difference map is displayed in the Figure while the real and imaginary parts are actually input to the networks.

computed by associating the detection of a moving area with the presence of more than 25 pixels with a score greater than 0.5 in the segmentation map. Conversely, for the classifiers, they are computed by thresholding the detection score by 0.5. It is worth mentioning that the training of the ResNet50 under a standard framework lasts 2h19 to compute 100 epochs with a NVIDIA Ampere A100 GPU. The training of the same network under the ECSPLAIN framework only took 10 more minutes to compute 100 epochs with the Grad-CAM, Layer-CAM and NormGrad. Since the use of Score-CAM is more memory demanding, the training is performed on a NVIDIA RTX A6000 and took 34 hours for 100 epochs.

6.4.3.2 Segmentation vs localization

- Is the localization task as powerful as the segmentation task?

Based on the performance of each network obtained in Table 6.5, the segmentation networks outperform the ECSPLAIN framework in terms of segmentation metrics. Yet, such a result was expected due to the larger number of parameters in these networks and their specificity to the segmentation task. Indeed, although the ResNet50 network used in the ECSPLAIN framework is trained to match its explainability with the segmentation map, its primary goal is image-wise labeling. Therefore, the main objective of this comparison is not to outperform the segmentation results - although the ECSPLAIN framework has done so in comparison to DeepLabV3 - but to evaluate whether the ECSPLAIN performance is comparable and competitive with that of the segmentation networks.

In particular, the performance of ECSPLAIN trained with NormGrad and Score-CAM ranks second and fourth respectively in terms of classification, with only 0.04 point difference in F1-Score between the first and fourth place. Looking at the precision and recall, the FCN metrics are more unbalanced than that of ECSPLAIN and the latter trained with the Score-CAM shows a balance similar to those of ResUNet. From a segmentation point of view, the framework trained with the Grad-CAM and the Score-CAM are highly competitive in terms of Dice and mIoU revealing their ability to appropriately localize the moving areas despite more uncertain maps as indicated by the threshold used to compute the metrics. In addition, ECSPLAIN with the Score-CAM achieves the best Hausdorff Distance of all and the second best Area Under the Curve, which is only 0.05 below the performance of ResUNet. Such performance can be better understood with the qualitative results in Figure 6.9. Indeed, as already experienced in Section 6.3.2, the segmentation networks suffer from under-segmentation of the moving areas. Such a behavior leads to the failure to detect certain moving regions, such as the top and bottom instabilities in the second row of Figure 6.9. Conversely, despite their coarser and more uncertain localization maps, ECSPLAIN trained with the Grad-CAM and the Score-CAM manages to highlight these particular regions while being accurate on the largest central movement. Therefore, the high performance in Hausdorff Distance is a direct consequence of the multi-object localization capability of the CAMs. For the same reason, and due to the coarseness of the localization maps, the MAE scores are significantly increased with the ECSPLAIN framework. Although this means that the localization maps are more uncertain than the segmentation maps, it also reveals that the

	Classification Metrics				Segmentation Metrics					Thresh.
	Acc. ↑	Prec. ↑	Rec. ↑	F1-S ↑	MAE ↓	Dice ↑	mIoU ↑	HD ↓	AUC ↑	
Segmentation Networks										
FCN	95.83%	**96.30%**	91.56%	0.939	**1.36e-2**	0.661	0.523	25.52	0.953	0.09
DeepLabV3	93.02%	86.99%	94.02%	0.904	1.64e-2	0.622	0.480	36.62	0.955	0.12
Res. U-Net	**95.96%**	93.89%	**94.55%**	**0.942**	1.43e-2	**0.684**	**0.547**	25.69	**0.964**	0.1
A posteriori explanation - ResNet50										
Grad-CAM	94.67%	92.43%	92.27%	0.923%	13.74e-2	0.161	0.093	58.01	0.524	0.10
Layer-CAM					13.00e2	0.350	0.230	45.95	0.813	0.26
NormGrad					13.76e-2	0.379	0.254	44.01	0.820	0.27
Score-CAM					14.20e-2	0.360	0.244	40.82	0.858	0.37
ECSPLAIN - ResNet50										
Grad-CAM	95.04%	92.66%	93.15%	0.929	3.35e-2	0.637	0.504	25.54	0.949	0.3
Layer-CAM	94.00%	94.02%	88.40%	0.911	4.72e-2	0.570	0.435	26.45	0.920	0.28
NormGrad	95.89%	95.63%	92.44%	0.940	4.73e-2	0.575	0.439	25.98	0.923	0.3
Score-CAM	95.71%	94.47%	93.16%	0.938	5.92e-2	0.643	0.500	**24.08**	0.959	0.33

Table 6.5: Quantitative comparison of the segmentation networks, *a posteriori* CAM localizations and ECSPLAIN framework. Both classification and segmentation metrics are provided. The last columns indicates the threshold used to binarize the localization maps and obtain the Dice, mIoU and Hausdorff Distance metrics. Best and second best results are bolded and underlined, respectively.

ECSPLAIN framework is able to measure the uncertainty and especially near the boundaries of the moving area according to Figure 6.9. This specific behavior is more in line with the expectations of a geoscientist looking for evidence of slope instability. In addition, although it may result in more false positive detections, this is highly preferable to the failure to detect actual phenomena as experienced with the segmentation networks in Figure 6.9. Therefore, despite the fact that ResUNet achieves the best quantitative results, the localization maps provided by the ECSPLAIN framework look more informative than the segmentation network outputs.

Based on the above segmentation-localization comparisons, the proposed framework is highly competitive with the results of the segmentation networks. In addition, the generated maps are more likely to provide clues about the actual presence or absence of moving areas in the region while gathering all the advantages mentioned in Section 6.4.1.

6.4.3.3 Explainability or ECSPLAIN-ability

- Is an explainability-constrained training actually more relevant than explaining a posteriori?

Despite the competitive results demonstrated by the ECSPLAIN framework compared to the segmentation networks, it does not prove its actual necessity. This section focuses on the evaluation of the relevance of the proposed framework compared to the application of CAMs *a posteriori* of the training, *i.e.* only at the inference time, as it is commonly used in the literature. The same Table 6.5 also summarizes the quantitative performance of the CAMs when applied *a posteriori* of the training while Figure 6.10 displays their qualitative performance compared to the ECSPLAIN framework.

From a segmentation point of view, an unquestionable conclusion can be drawn: the ECSPLAIN framework achieves a much better performance than the *a posteriori* CAMs. Since the network is trained to match the segmentation map, such a behavior could have been expected. Yet, the most striking point is the actual underperformance of the CAMs both quantitatively and qualitatively when they are applied *a posteriori*. Indeed, the Dice and mIoU scores do not reach 0.4 and 0.3, respectively, and the Hausdorff Distance is increased by almost 20 points compared to the ECSPLAIN framework results. These poor metrics are the direct consequence of the inability of the CAM to localize movements despite apparent motion patterns as illustrated in Figure 6.10. Indeed, on the localization maps generated by the *a posteriori* CAMs, the entire background is considered relevant, with little attention paid to the central apparent motion pattern. Only the Score-CAM is able to highlight a region that could be more convincingly associated with the moving area. Nevertheless none of the CAMs is able to localize the moving areas as accurately as their corresponding

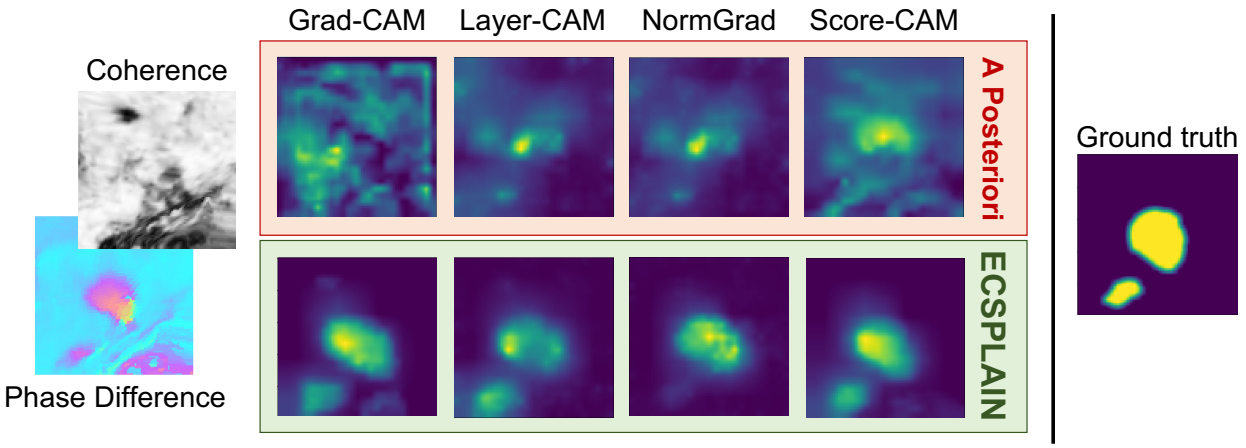


Figure 6.10: Qualitative comparison between the *a posteriori* CAM localization (top) and the ECSPLAIN framework (bottom). For visualization purpose, phase and coherence are displayed instead of the real and imaginary parts of the complex interferogram.

ECSPLAIN variant. Several factors may explain the poor performance of the CAMs applied at inference time, which then justifies that the network needs guidance to be reliable and explainable.

Interferogram characteristics: Unlike standard computer vision images, interferograms contain complex pixel values. Thus, peculiarities such as the rotation of the phase increase the difficulty of an already challenging task. Indeed, despite the fact that the interferogram is decomposed into its real and imaginary parts and that a data augmentation strategy is applied to increase the robustness of the networks, the classifiers have few parameters which limits their learning capabilities without external guidance.

Distortion patterns: In addition to the variety of patterns caused by moving areas on the phase difference map, the classifier should also avoid counterfactuals caused by distorted areas. Indeed, since slope instabilities are mainly located in mountainous regions, the interferograms are highly subject to distortions that can produce motion-like patterns (see Section 6.2.2). Moreover, moving area patterns are often located near distorted regions because once a slope is accurately imaged by the satellite, the backside slope is often distorted. Therefore, distinguishing the actual patterns of interest from the counterfactuals is a tough task for a lightweight network without external supervision. These limitations explain the inaccurate and highly uncertain localization maps generated by the CAMs and the significant decrease in the AUC score in Table 6.5.

Saliency maps: Most CAMs share the same initial goal: to localize the most salient features of the input images. By definition, such a behavior limits the detection of smaller patterns of interest with lower saliency but still related to the task without the use of complementary CAMs [Zhang et al., 2018b, Chen et al., 2022b, Qin et al., 2022]. The localization map obtained by Score-CAM at inference time in Figure 6.10 perfectly illustrates this issue as without the supervision provided by ECSPLAIN, only the largest motion pattern stands out from the uncertain background.

Selected layers: As reported by the authors of [Rebuffi et al., 2020], the choice of the layers is critical. In particular, they experienced that Grad-CAM shows poor localization ability when applied to the early layers. Therefore, even if the layers used for multi-level CAM generation are chosen to be those with the most meaningful activations, their relevance from a CAM to another may be questionable. This behavior is avoided in the ECSPLAIN framework since the explainability constraint forces each layer to be truly relevant not only for the localization task but also for the classification task.

As depicted by the above challenges, the guidance of the network is highly recommended to ensure the relevance, reliability and trustworthiness of the classifier detections. Indeed, based on the classification metrics given in Table 6.5, the F1-Score increases by up to 0.016 points when the network is guided. This particular data and the improvements in precision and recall also reveal that the guidance allows to limit the false positive detections due to the counterfactuals while not affecting the false negatives, and even re-

ducing them. Therefore, the proposed ECSPLAIN framework has three significant advantages over standard *a posteriori* CAM generation. First, it helps the network to find more relevant and discriminative patterns to improve the classification performance. Second, it ensures that the network is explainable and that the provided explanation is consistent with end-user expectations by generating the localization map. Finally, guidance is able to overcome, or at least mitigate, most of the above challenges to generate more accurate and insightful localization maps.

6.4.3.4 CAMs comparison

- Which CAM is most relevant for the localization of moving areas?

Based on the performance of the ECSPLAIN framework with the four CAMs in Table 6.5 and in the two Figures 6.9 and 6.10, each CAM has a different impact on the classifier network.

Indeed, from a strict classification point of view, NormGrad achieves the best results among the four CAMs in Table 6.5. Its ability to take into account more local features thanks to the pointwise gradient product seems to be appropriately leveraged by the network. The comparison of the performance of Layer-CAM and NormGrad also shows that the use of the Frobenius norm helps the network to better identify specific moving area features and thus to perform more accurate detections. Nevertheless, from a localization point of view and based on the segmentation metrics in Table 6.5, the ECSPLAIN frameworks trained with NormGrad and Layer-CAM show significant performance losses compared to Grad-CAM and Score-CAM. The reasons for such a quantitative counter-performance can be identified qualitatively in Figures 6.9 and 6.10. Indeed, in the former, similar to the segmentation networks, both Layer-CAM and NormGrad fail to detect the smallest moving areas in the second row and Layer-CAM even misses the left motion pattern in the first row. In addition, based on the second row and on the Figure 6.10, both NormGrad and Layer-CAM show less salient localizations compared to Grad-CAM and Score-CAM. Such a behavior is certainly caused by the nature of the target phenomenon. Indeed, moving areas are expected to be compact and to have a high spatial correlation. This characteristic favors the use of algorithms with more global receptive fields which can target the bulk motion patterns. Conversely, the use of local gradients for local CAM generation is particularly convenient for textured images where high-resolution features are the most discriminative. Therefore, by focusing on detecting specific high-resolution features on the main motion patterns, NormGrad and Layer-CAM cannot detect the presence of the smallest neighboring motion patterns.

Based on the performance of the two remaining CAMs, Grad-CAM and Score-CAM, in terms of classification and segmentation metrics in Table 6.5, the latter achieves the best balance. This is also visible in Figures 6.9 and 6.10 where the ECSPLAIN framework trained with Score-CAM shows higher saliency in the regions corresponding to moving areas. Yet, from a computational point of view, the use of Grad-CAM is more in line with the motivations described in Section 6.4.1. At inference time, the computation of Grad-CAM requires a single backward pass through the network to compute each of the gradient maps needed for the linear combination. Therefore, the number of FLOPs for detection and localization is doubled compared to the original ResNet50 FLOPs (see Table 6.3 in Section 6.4.1) which is still much better than segmentation networks. Nevertheless, the Score-CAM is significantly more expensive. Indeed, due to the fact that each masked image has to be passed to the network, and based on the number of kernels of each selected layer in Section 6.4.2, the number of FLOPs at inference time is at least multiplied by 1472. The slight performance gain of Score-CAM over Grad-CAM is not significant enough to justify such an increase in computation cost. Therefore, the most consistent method with respect to all the motivations detailed in Section 6.4.1 is the proposed ECSPLAIN framework trained with Grad-CAM as it shows the best balance in terms of classification-segmentation-computation metrics and quantitative-qualitative results.

6.4.3.5 CAM generalizability

- Is a network trained with a given CAM explainable with another CAM?

Since the ECSPLAIN framework trains a classifier to be explainable with a given CAM, special attention should be paid to the generalizability of the learned explainability. Indeed, as stated by most authors in

Inference CAM	MAE ↓	Dice ↑	mIoU ↑	H. Dist. ↓	AUC ↑
ECSPLAIN with Grad-CAM					
Grad-CAM	*3.35e-2*	*0.637*	*0.504*	*25.54*	*0.949*
Layer-CAM	15.95e-2	0.494	0.354	35.38	0.850
NormGrad	16.40e-2	**0.519**	0.378	**30.24**	0.865
Score-CAM	**14.91e-2**	0.518	**0.382**	32.85	**0.920**
ECSPLAIN with Layer-CAM					
Grad-CAM	**4.15e-2**	0.342	0.225	51.38	0.817
Layer-CAM	*4.71e-2*	*0.570*	*0.435*	*26.45*	*0.920*
NormGrad	4.95e-2	**0.562**	**0.429**	**26.70**	**0.918**
Score-CAM	9.00e-2	0.477	0.339	31.65	0.911
ECSPLAIN with NormGrad					
Grad-CAM	**3.88e-2**	0.317	0.206	47.66	0.787
Layer-CAM	4.64e-2	**0.579**	**0.439**	**26.11**	0.924
NormGrad	*4.73e-2*	*0.575*	*0.439*	*25.98*	*0.923*
Score-CAM	8.78e-2	0.485	0.346	39.63	**0.925**
ECSPLAIN with Score-CAM					
Grad-CAM	**9.23e-2**	0.244	0.152	56.58	0.727
Layer-CAM	11.31e-2	**0.529**	**0.385**	40.83	**0.890**
NormGrad	12.24e-2	0.519	0.379	**35.00**	0.881
Score-CAM	*5.92e-2*	*0.643*	*0.500*	*24.08*	*0.959*

Table 6.6: Quantitative comparison of the inference explanation generalizability of the ECSPLAIN framework depending on the CAM of training. Only segmentation metrics are provided. Best inference performance for each framework are bolded. Blue italic fonts refer to the reference inference where the same CAM is used for the training and the inference.

the CAM literature, the explanations provided by each CAM are not always consistent with the others [Rebuffi et al., 2020, Chattopadhyay et al., 2018]. Therefore, Table 6.6 gives the quantitative performance of each CAM applied at inference time on a ResNet50 classifier trained with another CAM in the ECSPLAIN framework.

The most striking conclusion from these results is the poor applicability and generalizability of Grad-CAM to the other CAMs. From an applicability point of view, the use of Grad-CAM at the inference time on the classifiers trained with another CAM shows the biggest performance degradation except for the MAE score. This counter performance can be explained by two factors. First, the Grad-CAM has poor localization capabilities in the early layers of the classifier as seen in [Rebuffi et al., 2020]. Indeed, since all CAMs are computed in multiple layers, this drawback implies a poor contribution of the high resolution feature maps to the localization task and thus a more uncertain CAM. This behavior is avoided when training the classifier with the ECSPLAIN framework with the Grad-CAM because the network is constrained to obtain an accurate localization map, which is not the case when the Grad-CAM is applied only at inference time. On the other hand, the Grad-CAM is generated with a global receptive field due to the average pooling of the gradient maps. Conversely, Layer-CAM and NormGrad have local receptive fields and Score-CAM is trained to leverage every particular feature resolution - including both global and local receptive fields. Therefore, at inference time, the computation of Grad-CAM cannot appropriately leverage the local features learned

by the other networks, resulting in a decrease in performance. Then, from a generalizability point of view, the performance of all CAMs for localizing the moving areas from the classifier trained with Grad-CAM is also greatly reduced. This is certainly due to the inverse of the latter receptive field problem: Grad-CAM provides global feature extraction that cannot be appropriately exploited by local CAM strategies.

Regarding the other CAMs, further considerations should be taken into account. First, Layer-CAM and NormGrad show very strong capabilities in terms of transferability from one to the other. Such a behavior is expected since the computation of their linear weights is similar, the main difference lies in the feature fusion strategy (linear combination against Frobenius norm, respectively). Based on Table 6.6, such a difference does not affect the performance of the one inferred on the classifier trained with the other. Then, Score-CAM shows the most balanced performance when inferred to the other classifiers. It shows that the masking approach has powerful capabilities in multi-resolution adaptability. In addition, the comparison of the performance of these three latter CAMs with that of the *a posteriori* CAMs in Table 6.5 further demonstrates the advantages of the ECSPLAIN framework in terms of generalizability of the explanation. Indeed, both the Dice and mIoU scores are increased by more than 0.1 points and the Hausdorff Distance is reduced by up to 15 points. Such performance supports the conclusions drawn in the previous discussion regarding the difficulty of the CAMs to explain the decision of the classifier without guidance.

6.5 ECSPLAINing ISSLIDE

This section aims to demonstrate the powerful capabilities of the ECSPLAIN framework described in Section 6.4. To this end, the framework trained with Grad-CAM is applied to find novel moving areas on the interferograms of the raw dataset of ISSLIDE detailed in Section 6.2. This application proposed in [Bralet et al., 2024a] allows to significantly enrich the number of susceptible moving areas in the dataset. Indeed, all the detected maps are included to the ISSLIDE dataset on IEEE Dataport[6] to be used as weak labels by future users.

6.5.1 Detection Pipeline

The entire detection pipeline is schemed in Figure 6.11 and can be divided into four different steps: one for the computation of the localization maps and three for the generation of the segmentation maps. The latter three are illustrated in Figure 6.11 with a patch extracted from the entire interferogram for better visualization, but these steps are actually performed on the entire interferograms.

Localization Map Generation: As already mentioned, this step consists in applying the ECSPLAIN framework to each of the interferograms of the raw ISSLIDE dataset. Nevertheless, due to the size of these interferograms and the limited memory resources, the entire interferogram cannot be processed at once. In addition, the ResNet50 architecture is based on a final Global Average Pooling layer which results in a significant loss of information when applied to the raw interferogram. Therefore, to avoid the aforementioned issues but also to be consistent with the training of the ECSPLAIN framework, a sliding window strategy is used to compute localization maps on patches of size 100×100px. To limit the impact of the sliding window and to ensure that all moving areas are fully visible on at least one patch, a stride of 25 pixels is used in both dimensions. Each patch is then fed to the ResNet50 classifier which was trained following the ECSPLAIN framework with Grad-CAM. The resulting overlapping localization patches are averaged to reconstruct the entire localization map.

Temporal Average: To limit the impact of false positives in the detections and to mimic the geomorphological guidelines given in Section 6.2.2, a temporal average is applied. To this end, all localization maps with the same temporal baseline are averaged pixel by pixel. Therefore, this particular step results in three different localization maps, one for each temporal baseline: 6, 12, and 18 days. The main drawback of this procedure is that it limits the detection of slow moving landslides which have longer periods of inactivity than rock glaciers.

[6] https://ieee-dataport.org/documents/isslide-insar-dataset-slow-sliding-area-detection-machine-learning, last visited on June 28, 2024.

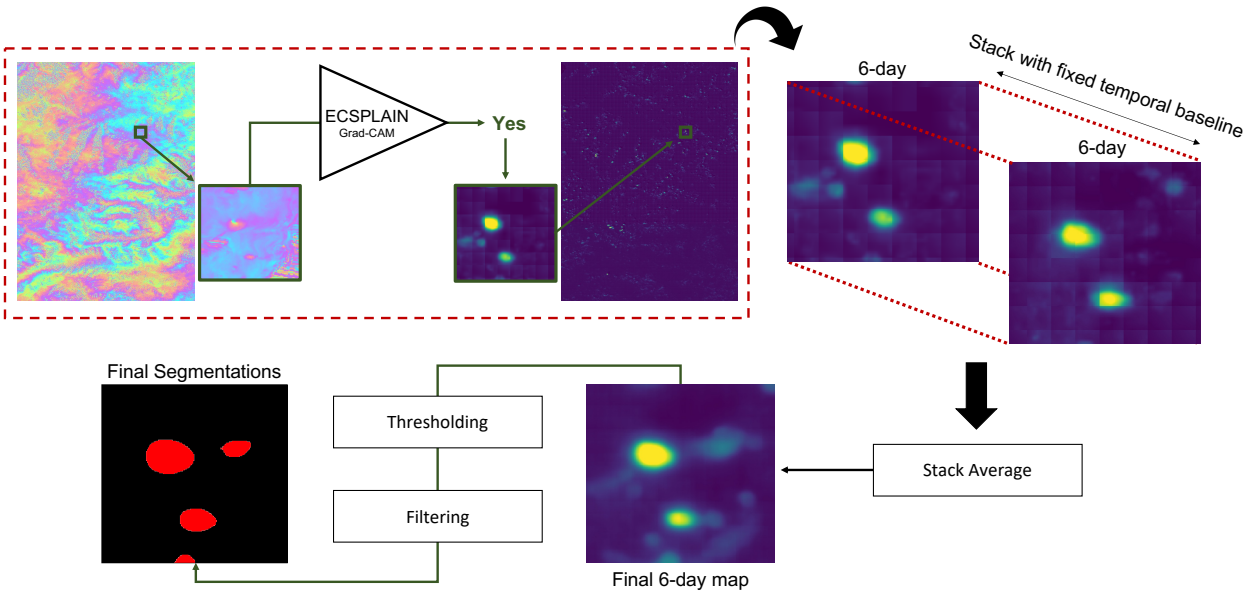


Figure 6.11: General pipeline for the generation of the moving area localization maps and their detected polygons over the whole interferograms. For the sake of readability, the scheme illustrates only the case of interferograms with a 6-days temporal baseline and a small crop is used for visualizing the post-sliding window process which is actually applied over the whole interferogram.

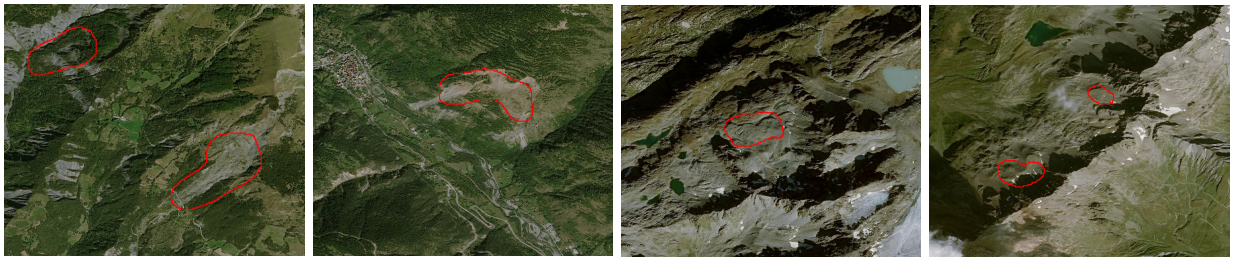
Map Binarization: The segmentation maps of the moving areas are then extracted from the temporal localization map by thresholding. According to the performance of ECSPLAIN trained with Grad-CAM in Table 6.5, the classifier achieves its most accurate segmentation performance with a threshold of 0.3. Yet, due to the slow moving landslide issue caused by the above temporal averaging and due to the susceptible lack of coherence caused by snowfall in high mountains or by sudden changes, a lower threshold is proposed. To ensure that the motion patterns are accurately detected (*i.e.* with a probability higher than 0.3) on at least 6 different interferograms with the same temporal baseline, the threshold is set to 0.1.

Moving Area Filtering: To ensure the consistency of the detections with the already annotated instabilities available in ISSLIDE, three different filters are applied to the resulting segmentation maps. First, all segmentations falling within an already annotated area are removed to avoid duplicates, assuming that the annotations are more reliable than the segmentations. Second, all false alarms detected within distorted areas or at the border of interferograms are automatically removed based on the layover and shadow maps used during the annotation process (see Section 6.2.2). Finally, to focus on the most likely moving areas and based on the statistics of the annotated areas, the last filter removes all the polygons with an area below $10,000\text{m}^2$.

Therefore, the pipeline results in two different products that are provided to the user as additional data in ISSLIDE. First, the localization maps computed for each of the interferograms of the raw dataset of ISSLIDE which can be used to extract further refined segmentation maps. Second, the shapefiles containing the delineated moving area polygons after the three post-processing steps which can be directly used for training deep learning networks.

6.5.2 Detected Instabilities

The application of the localization pipeline described above results in a number of novel moving area segmentations described in Table 6.7. These moving areas are classified in the table based on their size and the temporal baseline used for detection. According to these results, more than 400 moving areas are detected



(a) La Valette Landslide (b) La Clapière Landslide (c) Rock glacier near the Homme the Tachuy (d) Rock glaciers near the Col de la Tufe

Figure 6.12: Delineation of landslides and rock glaciers obtained by applying the segmentation pipeline based on the ECSPLAIN framework trained with Grad-CAM.

and localized for each temporal baseline. After removing the repetitiveness of the segmentations from one temporal baseline to another, a total of 831 susceptible new moving areas are identified by the proposed pipeline. Among them, 306 polygons are identified in both 6- and 12-day intergerograms, 194 in 6- and 18-day and 290 in 12- and 18-day. Finally, a total of 178 moving areas are common to the three temporal baselines. Based on the guidelines used during the annotation process in Section 6.2.2, the consistency of a moving area from one temporal baseline to another is an evidence of the presence of an actual motion in the region. Therefore, it not only illustrates the strong abilities of ECSPLAIN to be consistent from one interferogram to another and robust to different temporal baselines but also strengthens the confidence and relevance of the segmentations. Figure 6.12 illustrates some segmentations obtained with the pipeline. The two left images highlight the delineation (red) obtained over two well-known slow moving landslides, the La Valette and La Clapière landslides. On the other two images, three movements are detected near the Homme de Tachuy and Col de la Tufe. Based on the ridges and furrows visible on these optical images, these detections show the intrinsic characteristics of rock glaciers.

From a computational point of view, another key to assessing the relevance of the algorithm is the low false alarm rate obtained in the foreshortening areas. Indeed, during the distortion filtering stage of the pipeline, only 47, 30 and 27 polygons were removed in the 6-, 12- and 18-day inventories, respectively. Such a low number of false positives is an actual indicator that the network has learned to recognize and avoid counterfactual patterns and is robust to distorted areas.

Since these novel moving area segmentations have not yet been validated by geomorphological experts, they cannot be included as actual ground truth annotations in the ISSLIDE dataset. However, these data may be particularly useful for various deep learning approaches and strategies:

- **Fully supervised learning:** Based on the strong requirement of supervised learning approaches for confident and reliable ground truth, this additional data can be partially exploited. Indeed, strong evidence for accurate detection are: the size of the area, the temporal localization score and the temporal baseline redundancy. Therefore, a more reliable sub-dataset can be extracted by considering

	Number of Detections			
Area ($\times$ 1,000m^2)	[10,50]	[50,100]	>100	Total
6 days	308	118	89	515
12 days	282	121	92	495
18 days	252	96	82	430

Table 6.7: Statistics of the polygons discovered by our approach for each temporal baseline. Instabilities are clustered depending on their area ($\times 1,000m^2$) identified between brackets on top of columns.

only the polygons with an area greater than 50,000m^2, with a temporal consistency greater than 0.2 in the localization maps and/or which are common from one baseline to another. Since such a sub-dataset would be based on the most reliable segmentations provided by the pipeline, it could significantly improve the diversity of the dataset and thus the performance of deep neural networks.

- **Weakly supervised learning:** Since the localization maps are provided with the segmented polygons, they can be used as an evaluation of the reliability of the detections. Therefore, such information can be used to weight the loss and force the network to exploit the entire additional dataset of susceptible moving areas without full confidence.

- **Curriculum learning:** Such a strategy can be set up to exploit both the localization maps and all the segmentation maps. Indeed, while weakly supervised learning would use all the data to train the network from scratch, curriculum learning [Bengio et al., 2009] would progressively incorporate uncertain samples for training the network. Indeed, similar to the authors in [Zhang and Ma, 2021, Liu and Zhang, 2022], the simplest and most reliable segmentations can be used to train the network during the first epochs. Then from one epoch to another, more and more uncertain samples can be added to the dataset to train the network on a more complex and uncertain task while ensuring that the performance on easy samples is maintained.

Therefore, several strategies can be applied to this complementary dataset in order to appropriately exploit all the diversity of the detected phenomena and to make the networks even more robust and generalizable. The enrichment of the ISSLIDE dataset has the same objective as its creation: to encourage for the development and full exploitation of InSAR capabilities for the detection of ground deformation and, in particular, slope instability.

6.6 Conclusion

This final chapter focuses on the detection of slow moving slope instabilities. Unlike the previous Chapter 5, these instabilities are not immediately visible by their offset in image time series. Therefore, the phase of SAR images is exploited to identify slow motions of the order of centimeters. Through the computation of these so-called interferograms, a novel ISSLIDE dataset is proposed, which consists of manual annotations of slow moving landslides and rock glaciers. In particular, this chapter relies on the expert knowledge of geomorphologists to ensure accurate and precise identifications. The ISSLIDE dataset is then exploited to train standard deep neural networks designed for segmentation and to establish baseline performance. Finally, a novel ECSPLAIN framework is proposed to better meet the expectations of geoscientists. The latter is based on a standard classifier trained for the detection task. The peculiarity of the training lies in a new explainability constraint that forces the network to match its CAM with the annotation map of the underlying moving area. Such a strategy ensures that the network not only detects the presence of moving areas more accurately, but also computes a reliable localization map once a moving area is detected. The ECSPLAIN framework is further used to enrich the proposed ISSLIDE dataset with additional susceptible segmentations associated with a localization map that measures their reliability.

Based on the various results obtained throughout the chapter, the main conclusion can be summed up in two words: deeper research. Indeed, we have provided a novel dataset, shown that standard segmentation approaches are efficient yet imperfect, and implemented a very promising explainability strategy that takes into account both of the geoscience and computer science worlds. All these results and tools pave the way for novel deep learning approaches applied to the detection and monitoring of slope instabilities before they fail. In addition, considering interferograms computed on higher resolution SAR images would certainly allow for even more precise detection and localization capabilities. Such research would therefore certainly contribute to the mitigation of the threats posed by these hazards.

Related communications

Journal

[Bralet *et al.*, GRSL 2024] *ISSLIDE: A new InSAR dataset for Slow SLIding area DEtection with machine learning* , Bralet, A., Trouvé, E, Chanussot, J., Atto, A. M., in IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024, Art no. 3001005, DOI: 10.1109/LGRS.2024.3365299.

[Bralet *et al.*, TGRS 2024] *ECSPLAIN: Explainability Constrained-claSsifier for Pairing detection and Localization of moving Areas on INterferograms*, Bralet, A., Atto, A. M., Chanussot, J., Trouvé, E., in IEEE Transactions on Geoscience and Remote Sensing - submitted on June 10, 2024.

Dataset

[Bralet *et al.*, IEEE Dataport 2023] *ISSLIDE: InSAR dataset for Slow SLIding area DEtection with machine learning*, Bralet, A., Trouvé, E., Chanussot, J., Atto, A. M., on IEEE Dataport, September 22, 2023, doi: https://dx.doi.org/10.21227/dhxt-5g91.

Chapter 7

Conclusion and Perspectives

Slope instabilities, medium resolution remote sensing images and deep learning are three different topics that may sound incompatible from some points of view, but which worked in harmony in the thesis. To this end, a multitude of difficulties have been overcome. From geometrical distortions to cloudy images, from relevance to credibility, from lack of datasets to geomorphological evidence, from complex values to understandability, so many challenges that motivated the contributions proposed in this thesis (and the PhD student who I am). First, this chapter proposes a summary of these contributions to give a better idea of the evolution of the thoughts, strategies and solutions investigated in the thesis. Three main research areas have been targeted in this work: modality translation, slope instability change detection on bitemporal images and slope instability inventory on interferograms. Then, the chapter ends with further research directions and perspectives that we propose based on the conclusions resulting from the studies of the thesis.

7.1 Modality Translation *- Clouds or geometrical distortions? That is the question!*

The SAR-to-optical modality translation is introduced in this thesis as a promising approach to address two major issues related to the combination of remote sensing and slope instabilities: the cloud cover induced by rainfall events and the geometrical distortions caused by slopes. By translating a SAR image into an optical image, the cloudy optical images can be substituted by the translated images ensuring the availability of an image at any time. In addition, an accurate translation should transfer the actual structural properties of the SAR image while mitigating the effects of geometrical distortions.

The first contribution proposed in the thesis is the implementation of the SAR Distorted Image translator Network (SARDINet). This neural network constructed under an encoder-decoder architecture successfully identifies the structural features in an urban area and translates them into relevant optical patterns. SARDINet even shows more robust performance against the layover and shadow distortions than standard but widely explored adversarial networks from the literature.

The architecture is further improved to better meet the requirements in rural areas and especially in mountainous regions with steep slopes. To this end, a first improvement is the integration of the Digital Elevation Model (DEM) as complementary information. Since the DEM is correlated with the slope which causes geometrical distortions in SAR images, it seems appropriate to include it as an input to the network. Therefore, a novel dataset in mountainous regions containing triplets of DEM-SAR-optical images has been made available[1] for further works on the geometrical distortion issues. Different fusion strategies are then implemented to combine the DEM and SAR images. The implementations end with similar conclusions: the DEM is a valuable information to improve the relevance of the optical patterns to be reconstructed, *e.g.* by

[1] https://ieee-dataport.org/documents/sar-dem-optical-mountainous-dataset-distortion-management, last visited on June 28, 2024.

reducing the confusion between water bodies and shadow areas. The decoding layers used to upsample the low-resolution latent code of SARDINet are also fine-tuned by comparing five different strategies. Among the conclusions drawn by the study, the transpose convolutions succeed in reconstructing the most relevant hidden features while the use of a sub-pixel convolution layer [Shi et al., 2015] at the end of the network allows a better reconstruction of high-resolution features.

After the significant increase in pattern relevance, further works are proposed to increase the credibility of the reconstructed images without affecting their relevance. Three different suggestions are implemented:

- SARDINetV2 with architectural modifications such as skip connections, latent residual blocks and learnt branch fusion.
- Adv-SARDINet with an auxiliary discriminator network to train SARDINet adversarially,
- Standard post-processing applied after the translation to improve the contrast of the image.

SARDINetV2 shows the most interesting results of all, as it successfully improves the relevance of the reconstructed patterns while increasing the contrast without any additional processing. Given the pros and cons of each method, a Translation-Deblurring (Tra-De) framework is still under experimentation to apply a standard image-to-image adversarial network (*e.g.* Pix2pix [Isola et al., 2017]) to enhance the contrast of the translated images from SARDINetV2. However, at this point, although the Tra-De framework significantly improves the credibility of the image, it still suffers from the drawbacks of adversarial networks. First, the instability of the training makes it difficult to train and fine-tune to ensure accurate results. Second, the credibility of the results is traded off against the relevance of the reconstructed patterns. Nevertheless, the translated images are intended to replace the cloudy optical images to provide early slope instability collapse detection. Therefore, the deterioration of the relevance of the patterns makes the application of the Tra-De framework to this task very limited for now.

7.2 Bitemporal Change Detection

- Spot the landslides!

After the implementation of the SAR-to-optical modality translation, the thesis addresses the challenge of detecting slope instabilities. Since these events are punctual in time and space and poorly reported, most of the datasets available for SAR-to-optical translation do not include slope instability samples - and in particular landslide collapses. Therefore, even if a translator achieves accurate results on the datasets from the literature, applying its knowledge to reconstruct landslide patterns is another key challenge of the thesis.

A first standard strategy, which has already shown strong performance in computer vision and remote sensing, is a simple fine-tuning of the network on the target dataset. In the case of the present study, it consists in retraining SARDINetV2 for a few epochs with the new proposed dataset composed of pre- and post-landslide images. To this end, another dataset is made available[2] for the detection of landslide collapse on bitemporal SAR and optical images. Nevertheless, the application of change detection networks to the translations of SARDINetV2 fine-tuned on such a dataset did not achieve the expected performance. Among the reasons for such a counter performance, two are further addressed in the thesis:

- Even if the translated images are relevant from a *metric* point of view, this does not mean that they are relevant from an *application* point of view.
- The translated image is not similar enough to the pre-event optical image, which limits the possibilities for comparison.

Therefore, we propose to integrate the knowledge of a downstream task in the training of our SAR-to-optical translator network. To better evaluate these intuitions, a first Translation-Classification Loss (TCL) is proposed to optimize the training of the translator on BigEarthNet-MM [Sumbul et al., 2021] which contains

[2]https://ieee-dataport.org/documents/multimodal-remote-sensing-dataset-landslide-change-detection-haiti, last visited on Jue 28, 2024.

a large number of multi-labeled images. The TCL consists in leveraging a pre-trained and frozen multi-label classifier to evaluate the relevance of the translated image. Since the classifier is trained on real optical images and frozen during the training of the translator, it expects the translator to reconstruct credible features with respect to its task of interest. Thus, instead of adversarially training the network with a discriminator, the TCL proposes to use the auxiliary classifier as a guide to increase the relevance of the reconstructed patterns and make them even more specific to the downstream task. The TCL has shown strong improvements not only on the translations of the proposed SARDINet and SARDINetV2 architectures, but also on the results of four other standard adversarial approaches from the literature by substituting the discriminator with the frozen classifier. Although the credibility of the images is reduced due to the substitution, a strong increase in pattern relevance is observed. Moreover, when both the TCL and an adversarial training are coupled, the former shows a strong regularization ability and limits the generation of artifacts caused by the use of the discriminator.

Encouraged by the performance of the TCL, a Translation-Change Detection Loss (TCDL) is being experimented to specialize the translation for landslide detection on bitemporal images. However, due to the early experimental stage of the TCDL, no clear and final conclusions can be drawn at this time. From a translation point of view, the TCDL seems to better guide SARDINetV2 to reconstruct more accurate landslide patterns. However, the application of change detection networks still leads to inaccurate detections. Further and deeper optimizations and architectural changes should be conducted to better guide the translations in reconstructing change detection relevant patterns. More details on the further works to be investigated are given in the following Section 7.4.

7.3 Interferogram Inventorying - *A game of cats and slope instabilities*

Since this thesis is concerned with the detection of both sudden and slow moving slope instabilities, the last set of contributions focuses on the latter. The identification of slope instabilities before a collapse is always a challenging task, since it involves the search for precursors to mitigate the threat before the collapse. In this thesis, we have considered a step prior to the identification of collapse precursors: the inventory of these slope instabilities.

However, because slow moving slope instabilities show displacements on the centimeter scale, they are poorly visible, if not invisible, in medium-resolution satellite optical images such as Sentinel-2. This issue is addressed by considering SAR interferometry (InSAR) which leverages the phase of the radar images to reveal centimeter-scale displacements between two images. Due to the lack of datasets in the literature considering such application and data, a novel Interferogram dataset for Slow SLIding area Detection (ISSLIDE) is proposed. First, a standard and generalizable routine is applied to generate the interferograms used for manual annotation of moving areas in the French Alps. Following guidelines provided by geomorphological experts, 200 independent moving areas are manually identified in each of the 54 generated interferograms. The final ISSLIDE dataset is made available on IEEE Dataport[3] in two different configurations: a raw dataset for enrichment and custom instability extraction, and a Ready to Be Used (R2BU) dataset already adapted for training deep learning models. The dataset is particularly rich as it allows multi-temporal and multi-baseline approaches, making it applicable for both inventory and monitoring purposes.

Standard segmentation networks are applied to the ISSLIDE dataset to identify the limitations of such approaches due to the nature of the data and the studied phenomena. The results show that although the performance of these networks remains correct in terms of detection accuracy, they lack explanations and flexibility with respect to user requirements. In addition, they demonstrate that the knowledge acquired by the networks trained on standard computer vision images is not transferable to interferograms, requiring a retraining of the network from scratch. Finally, from a computer science point of view, such heavy networks are demanding in terms of memory and floating-point operations.

To mitigate the issues observed with standard segmentation networks, a novel Explainability-Constrained claSsifier for Pairing detection and Localization of moving Areas on INterfrograms (ECSPLAIN) has been

[3]https://ieee-dataport.org/documents/isslide-insar-dataset-slow-sliding-area-detection-machine-learning, last visited on July 15, 2024

proposed. The framework proposes to pair the detection and localization tasks based on a simple classifier network that is trained to accurately explain its decisions. To this end, Class Activation Maps (CAM) are computed during the training of the classifier. The latter is optimized for its ability to accurately detect the presence or absence of moving areas on the input interferogram patches but also for the relevance of the generated CAMs with respect to the true annotation maps. In this way, the explainability of the network is used to support the decisions of the network and to provide an additional informative output to the end user: the localization map. Different CAMs have been used for the training to better evaluate their strengths and weaknesses, and Grad-CAM is identified as a suitable CAM candidate for general use. Indeed, it accurately identifies the localization of slope instabilities while maintaining a low computational cost at inference time. For these reasons, it has been applied to enrich the ISSLIDE dataset based on its predictions computed on the entire interferograms. It should be noted that Score-CAM achieves the most accurate results among the CAMs but due to its longer and heavier computation, it is less relevant for early detection purposes.

7.4 Gates Are Opened!

– The best is the enemy of the good!... but still...

According to the different conclusions drawn in this thesis, the encouraging preliminary results, and the existing literature in computer vision and remote sensing, one cannot consider this work as an actual end. Therefore, this section proposes several promising research directions derived from the contributions and studies explored in the thesis.

7.4.1 Towards more Relevant Translations

From the point of view of geometrical distortion-aware modality translation, several directions can be explored, as it is a complex and ill-posed task. Nevertheless, it is worth mentioning that the focus of this thesis was to ensure the relevance of the patterns of the translated images. Therefore, the credibility of the patterns should not prevail on their relevance, *i.e.* relevant but blurry reconstructions are considered more interesting than attractive images that contain artifacts. The different research directions proposed here keep the same spirit.

1. ***Focus on non-adversarial translations:*** Due to their high instability and their tendency to hallucinate artifacts, adversarial networks should be considered with strong precautions for SAR-to-optical translation in areas affected by geometrical distortions. Although the blurry aspect of the generator-only neural networks may not look natural, they ensure the reconstruction of relevant patterns associated with the structural features of the SAR image. In addition, due to the lack of information in distorted areas, an adversarial network would be forced to hallucinate patterns without any indication of their actual presence. Conversely, the generator-only architecture will only guess the most susceptible type of land cover but without generating unreliable features which avoids misguidance of downstream applications. Therefore, generator-only networks are mostly encouraged for reliable translation.

2. ***Integrate more external prior information:*** One of the main reasons for blurry translations is the lack of spectral information in the SAR image. Therefore, a first idea could be to leverage a previously available optical image of the underlying terrain to reconstruct even more relevant translations. The main risk of using such a past image is its dominant use by the network. Indeed, since the images are taken over the same area, a simple solution for translating the SAR image into an optical image is to copy the provided past optical image. However, if a change has occurred between the two dates, it could be missed in the reconstruction since the focus would be on the pre-event image. Another interesting prior information lies in the data related to the underlying terrain. Indeed, only the Digital Elevation Model (DEM) is considered in this work but the slope, the aspect, the layover and shadow masks, and even the land-cover map could be valuable information to input into the network. Such maps could help to better identify areas suffering from geometrical distortions as well as the nature and composition of the landscape.

3. ***Better integrate prior information:*** As mentioned above and observed in the thesis, one of the main issues with multi-input networks is their appropriate use by the network. The dominance of one input over the other can quickly arise which would limit the use of the other inputs. Therefore, more attention should be paid to the integration of the different inputs so that they can help to exploit the SAR image features and not replace them. As experienced in the thesis, the late fusion strategy may be more prone to extract input-specific features, but their integration may still be questionable. To avoid such troubles, a promising concept is to condition the network with alternative data. In particular, the Conditioned Batch Normalization proposed in [de Vries et al., 2017, Dumoulin et al., 2017, Mescheder et al., 2019] would only affect the normalization step of the features extracted from the SAR images, forcing the network to focus on SAR feature extraction while using other data as advisors. Another possibility is to use attention modules to generate attention maps based on the auxiliary data and help to focus on more specific SAR features during the training, while providing a kind of "learned explainability".

4. ***Regularize post-deblurring networks:*** Finally, if the credibility of the generated patterns can be considered, it should not reduce their relevance. The preliminary results obtained with the Tra-De framework are promising in terms of contrast enhancement despite the decrease in pattern relevance. Thus, more attention should be paid to the regularization of the deblurring stage in order to limit the artifact hallucinations caused by the adversarial training used for texture generation. Such regularization can be related to the calculation of the loss as experienced with TCL or the design of the discriminator [Zhang et al., 2024].

7.4.2 Towards more SAR-optical Feature Comparisons

Therefore, several improvements can still be considered in standard modality translation. Yet, the proposed improvements do not specifically force the network to meet the requirements of a given downstream task. In the case of change detection, high expectations are placed on the comparison of the features obtained from the pre- and post-event images. Yet, if the distribution of each image is too different from the other, they can result in multiple inaccurate detections. To address this issue, the focus can be on adapting the translation task or modifying the whole strategy based on the literature.

1. ***Focus on collaborative training:*** In the case of sudden landslide detection, the translations of SARDINetV2 were not specific enough to ensure accurate performance compared to the monomodal approaches. Therefore, in this work, a downstream task network is used to reconstruct task-related features. Based on the results of the TCL and the preliminary results of the TCDL, using the knowledge of a pre-trained downstream task to improve the relevance of translation shows promising performance. The collaboration between the two tasks ensures the reconstruction of accurate patterns which can then be appropriately used by the downstream network, either directly or after retraining a novel network to tackle the downstream task with the translated images.

2. ***Progressively train the network:*** Another issue related to the dataset used to train the change detection network in this thesis is the cloud cover on the pre-event images. To address this problematic, one solution could be to implement a curriculum learning approach [Bengio et al., 2009] that would first train the network on cloud-free images and progressively include more cloudy images in the training dataset so that the network first focuses on the relevant patterns and then learns to consider clouds as irrelevant patterns.

3. ***Extract missing modality features:*** An approach similar to image translation is the extraction of features related to a missing modality from another available modality. Staying in the SAR-optical context where the optical image is missing, the hallucination networks proposed in [Hoffman et al., 2016] consist in extracting optical specific features from the SAR image using an auxiliary feature extractor. Such a strategy may be less demanding than modality translation networks since the ill-posed problem caused by the reconstruction of multispectral bands from the SAR image disappears.

Yet, the challenge remains complex since it consists of a low-level translation: extracting the structural properties of the SAR image and translating them into optical specific features. The comparison of the latent features can then be used to highlight the changes that have occurred between the acquisitions.

4. ***Implement representation learning:*** Another strategy derived from the previous one is to force the matching of both SAR and optical feature spaces. Instead of translating SAR features into optical ones, the idea is to find a common representation of these two images in a given latent space. Several works already focus on such a strategy [Luppino et al., 2024, Jain et al., 2021], but the identification of changes remains challenging, especially when a specific category of changes is targeted, here slope instability failures.

7.4.3 Towards more Geophysically-based Detections

The final set of contributions proposed in this thesis definitely paves the way for novel approaches that could combine the powerful generalization and computational capabilities of deep learning networks with physical and expert expectations. Not only the provided ISSLIDE dataset is based on a geomorphologically consistent routine, but the ECSPLAIN framework proposes to enforce the explainability of the network to make the decisions more trustworthy. Based on these results and the approaches already experimented, the interactions between geoscience and deep learning can be further strengthened by exploring many research directions.

1. ***Apply multi-temporal approaches:*** Due to the high revisit frequency of satellites, the development of multi-temporal approaches is a core aspect of remote sensing imagery. Although the thesis focused mainly on bitemporal approaches through change detection and interferograms, the use of a time series is the logical extension of these works. From a change detection point of view, the detection of landslides would appear as a severe break in the continuity of pixel values. However, further challenges should be addressed with respect to all the other types of changes that can occur, and in particular crop harvesting since a similar behavior appears: vegetation is suddenly switched to rocks and earth patterns. From the point of view of the detection of slow moving areas, the annotation routine developed together with geomorphologists insists on the importance of the redundancy of the movement to ensure the actual presence of a slope instability. Therefore, the inclusion of temporality in the detection and localization strategies will more closely mimic the actual expertise of geomorphologists.

2. ***Apply multi-baseline approaches:*** Similarly, the geomorphological routine encourages to consider the detected motion under different temporal baselines for even more reliable results. Therefore, the implementation of approaches which take as input a stack of interferograms with multiple temporal baselines would certainly help to better identify the phenomena. The most challenging issue with both multi-temporal and multi-baseline approaches is the necessary increase in complexity of the neural networks. Nevertheless, lightweight networks have shown promising results throughout the thesis. Such approaches could be further considered to reduce the complexity of the networks while staying as close as possible to the geomorphological guidelines.

3. ***Integrate more physics-based explanations:*** In the current stage of the thesis, the explanation used for the training is provided at the image level. However, a number of physical properties can be evaluated from sudden changes and from interferograms. One sophisticated research direction could be the implementation of Physics-Informed Neural Networks (PINN) [Raissi et al., 2019]. Coupling the explanations together with physical parameters would certainly increase the appeal of deep neural networks in the geoscience community. Assessing the area, volume or even estimating the damage caused by sudden landslides simultaneously with their detection could be informative enough to measure the level of alarm of the threat. Similarly, when working on interferograms, the estimation of the displacement in the Line Of Sight (LOS) but also the actual susceptible displacement based on the slopes and aspect of the underlying area can better inform on the state of activity of the hazard.

4. ***Implement monitoring and early warning methods:*** The end point of the three previous directions could be the implementation of lightweight networks capable of monitoring the evolution of the landscape. From an early monitoring point of view, such strategies could help to follow the evolution of any moving areas on a large scale in order to better inform local authorities about the subsurface threats. From an early warning point of view, the network could be responsible for detecting the regions that need special attention before a collapse, *i.e.* when an acceleration of the displacement is found out. From a post-event monitoring point of view, based on the detection of collapses, appropriate decisions can be made regarding the state of the hazard and whether it has turned into a disaster or not. Not only could the trustworthiness of the algorithm be ensured through physical parameter estimation and network explainability strategies, but the generalizability and computational efficiency of deep learning could also be leveraged for faster decision making.

7.5 One Last Word *– Is it really the last?*

Looking back at the conclusions drawn in this chapter, as well as the various contributions and challenges faced in the previous chapters, few corrections should be made.

I feel like a liar when I answer *"Basically, I ask my computer to look for slope instability-related phenomena on remote sensing images"* to the question *"What do you do?"*. Indeed, throughout this thesis, we have found several conclusions and promising research directions to improve the detection accuracy and hopefully reduce the impact of slope instabilities. We addressed a number of challenges related to deep learning, remote sensing and slope instability research areas. And we created more interactions and collaborations between computer science and geoscience communities.

For all of these reasons, I may change my answer to better figure out that during my PhD,

I faced the challenges associated with stubborn learning algorithms applied to remote sensing images that are more than "images" to prevent donkeys from sliding into your living room, since implementing studious algorithms is my way of keeping the donkeys away.

Bibliography

[Adriano et al., 2020] Adriano, B., Yokoya, N., Xia, J., Miura, H., Liu, W., Matsuoka, M., and Koshimura, S. (2020). Learning from Multimodal and Multitemporal Earth Observation Data for Building Damage Mapping. *ISPRS J. Photogramm. Remote Sens.*, 175:132–143.

[Aitken et al., 2017] Aitken, A., Ledig, C., Theis, L., Caballero, J., Wang, Z., and Shi, W. (2017). Checkerboard Artifact Free Sub-pixel Convolution: A Note on Sub-pixel Convolution, Resize Convolution and Convolution Resize. arXiv:1707.02937 [cs].

[Amankwah et al., 2022] Amankwah, S. O. Y., Wang, G., Gnyawali, K., Hagan, D. F. T., Sarfo, I., Zhen, D., Nooni, I. K., Ullah, W., and Zheng, D. (2022). Landslide Detection from Bitemporal Satellite Imagery Using Attention-Based Deep Neural Networks. *Landslides*.

[An et al., 2016] An, H., Viet, T. T., Lee, G., Kim, Y., Kim, M., Noh, S., and Noh, J. (2016). Development of Time-Variant Landslide-Prediction Software Considering Three-Dimensional Subsurface Unsaturated Flow. *Environ. Model. Softw.*, 85:172–183.

[Anantrasirichai et al., 2019] Anantrasirichai, N., Biggs, J., Albino, F., and Bull, D. (2019). The Application of Convolutional Neural Networks to Detect Slow, Sustained Deformation in InSAR Time Series. *Geophys. Res. Lett.*, 46(21):11850–11858.

[Anantrasirichai et al., 2018] Anantrasirichai, N., Biggs, J., Albino, F., Hill, P., and Bull, D. (2018). Application of Machine Learning to Classification of Volcanic Deformation in Routinely Generated InSAR Data. *J. Geophys. Res. Solid Earth*, 123(8):6592–6606.

[Aslam et al., 2021] Aslam, B., Zafar, A., and Khalil, U. (2021). Development of Integrated Deep Learning and Machine Learning Algorithm for the Assessment of Landslide Hazard Potential. *Soft. Comput.*, 25:13493–13512.

[Aslan et al., 2020] Aslan, G., Foumelis, M., Raucoules, D., De Michele, M., Bernardie, S., and Cakir, Z. (2020). Landslide Mapping and Monitoring Using Persistent Scatterer Interferometry (PSI) Technique in the French Alps. *Remote Sens.*, 12(8):1305.

[Bai et al., 2024] Bai, X., Pu, X., and Xu, F. (2024). Conditional Diffusion for SAR to Optical Image Translation. *IEEE Geosci. Remote Sens. Lett.*, 21:1–5.

[Batson and Royer, 2019] Batson, J. and Royer, L. (2019). Noise2Self: Blind Denoising by Self-Supervision. In *International Conference on Machine Learning*, pages 524–533, Long Beach, CA, USA. PMLR.

[Bayer et al., 2017] Bayer, B., Simoni, A., Schmidt, D., and Bertello, L. (2017). Using Advanced InSAR Techniques to Monitor Landslide Deformations Induced by Tunneling in the Northern Apennines, Italy. *Eng. Geol.*, 226:20–32.

[Benedetti et al., 2018] Benedetti, P., Ienco, D., Gaetano, R., Osé, K., Pensa, R., and Dupuy, S. (2018). M^ 3\text {Fusion} $: A Deep Learning Architecture for Multiscale Multimodal Multitemporal Satellite Data Fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 11(12):4939–4949.

[Bengio et al., 2009] Bengio, Y., Louradour, J., Collobert, R., and Weston, J. (2009). Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 41–48, New York, NY, USA. Association for Computing Machinery.

[Bermudez et al., 2018] Bermudez, J. D., Happ, P. N., Oliveira, D. a. B., and Feitosa, R. Q. (2018). SAR to Optical Image Synthesis for Cloud Removal with Generative Adversarial Networks. *ISPRS Ann. Photogramm. Remote Sens.*, IV-1:5–11. ISSN: 2194-9042.

[Bertone et al., 2022] Bertone, A., Barboux, C., Bodin, X., Bolch, T., Brardinoni, F., Caduff, R., Christiansen, H., Darrow, M., Delaloye, R., Etzelmüller, B., Humlum, O., Lambiel, C., Lilleøren, K., Mair, V., Pellegrinon, G., Rouyet, L., Ruiz, L., and Strozzi, T. (2022). Incorporating InSAR Kinematics into Rock Glacier Inventories: Insights from 11 Regions Worldwide. *The Cryosphere*, 16(7):2769–2792.

[Biggs et al., 2021] Biggs, J., Dogru, F., Dagliyar, A., Albino, F., Yip, S., Brown, S., Anantrasirichai, N., and Atıcı, G. (2021). Baseline Monitoring of Volcanic Regions with Little Recent Activity: Application of Sentinel-1 InSAR to Turkish Volcanoes. *J. Appl. Volcanol.*, 10(1):1–14.

[Bodin, 2020] Bodin, X. (2020). Rock glacier inventory of the Vanoise massif (French Alps). Research Report, EDYTEM, Université de Savoie Mont-Blanc, UMR 5204 CNRS.

[Bodin et al., 2018] Bodin, X., Echelard, T., Lacroix, P., and Strozzi, T. (2018). Regional Evaluation of Rock Glacier Activity in the Semi-Arid Andes Using Optical and Radar Satellite Imagery. In Edited by Philip Deline, X., Ravanel, L., Barboux, C., Delaloye, R., Lambiel, C., Magnin, F., Marcer, M., Pogliotti, P., and Schoeneich, P., editors, *5 th European Conference On Permafrost*, 5 th European Conference On Permafrost – Book of Abstracts,, Chamonix, France. Edytem.

[Bodin et al., 2017] Bodin, X., Krysiecki, J.-M., Schoeneich, P., Le Roux, O., Lorier, L., Echelard, T., Peyron, M., and Walpersdorf, A. (2017). The 2006 Collapse of the Bérard Rock Glacier (Southern French Alps). *Permafr. Periglac. Process.*, 28(1):209–223. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ppp.1887.

[Bonì et al., 2018] Bonì, R., Bordoni, M., Colombo, A., Lanteri, L., and Meisina, C. (2018). Landslide State of Activity Maps by Combining Multi-Temporal A-DinSAR (LAMBDA). *Remote Sens. Environ.*, 217:172–190.

[Bountos et al., 2022a] Bountos, N., Papoutsis, I., Michail, D., and Anantrasirichai, N. (2022a). Self-Supervised Contrastive Learning for Volcanic Unrest Detection. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Bountos et al., 2022b] Bountos, N., Papoutsis, I., Michail, D., Karavias, A., Elias, P., and Parcharidis, I. (2022b). Hephaestus: A Large Scale Multitask Dataset Towards InSAR Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1453–1462, New Orleans, LA, USA.

[Boyd et al., 2023] Boyd, A., Tinsley, P., Bowyer, K. W., and Czajka, A. (2023). CYBORG: Blending Human Saliency Into the Loss Improves Deep Learning-Based Synthetic Face Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6108–6117, Waikoloa, HI, USA.

[Bralet et al., 2022] Bralet, A., Atto, A. M., Chanussot, J., and TrouvÉ, E. (2022). Deep Learning of Radiometrical and Geometrical Sar Distorsions for Image Modality Translations. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1766–1770, Bordeaux, France.

[Bralet et al., 2023] Bralet, A., Atto, A. M., Chanussot, J., and Trouvé, E. (2023). Impact De La Stratégie De Décodage Sur La Traduction De Modalité Radar-Optique D'images De Télédétection. In *GRETSI 2023 - XXIXème Colloque Francophone de Traitement du Signal et des Images*, volume 2023, pages 929–932, Grenoble, France. GRETSI - Groupe de Recherche en Traitement du Signal et des Images.

[Bralet et al., 2024a] Bralet, A., Atto, A. M., Chanussot, J., and Trouvé, E. (2024a). ECSPLAIN: Explainability Constrained-claSSifier for Pairing detection and Localization of moving Areas on INterferograms. *IEEE Trans. Geosci. Remote Sens.* submitted on June 10, 2024.

[Bralet et al., 2024b] Bralet, A., Atto, A. M., Chanussot, J., and Trouvé, E. (2024b). Translation-classification loss for SAR image understanding with deep learning. *Comput. Vis. Image Underst.* submitted on April 19, 2024.

[Bralet et al., 2024c] Bralet, A., Ngo, T. N., Trouvé, E., Chanussot, J., and Atto, A. M. (2024c). DEM-assisted neural network for SAR-to-optical image translation. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, Athens, Greece.

[Bralet et al., 2024d] Bralet, A., Trouvé, E., Chanussot, J., and Atto, A. M. (2024d). ISSLIDE: A New InSAR Dataset for Slow SLIding Area DEtection With Machine Learning. *IEEE Geosci. Remote Sens. Lett.*, 21:1–5.

[Brengman and Barnhart, 2021] Brengman, C. and Barnhart, W. (2021). Identification of Surface Deformation in InSAR Using Machine Learning. *Geochem. Geophys. Geosyst.*, 22(3):e2020GC009204.

[Brideau and Roberts, 2022] Brideau, M.-A. and Roberts, N. J. (2022). Chapter 3 - Landslides in Bedrock. In Davies, T., Rosser, N., and Shroder, J. F., editors, *Landslide Hazards, Risks, and Disasters (Second Edition)*, Hazards and Disasters Series, pages 43–97. Elsevier.

[Bromley et al., 1993] Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., and Shah, R. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. In *Advances in Neural Information Processing Systems*, volume 6, pages 1–8, Denvers, CO, USA. Morgan-Kaufmann.

[Burrows et al., 2019] Burrows, K., Walters, R., Milledge, D., Spaans, K., and Densmore, A. (2019). A New Method for Large-Scale Landslide Classification from Satellite Radar. *Remote Sens.*, 11(3):237.

[Cambrin et al., 2023] Cambrin, D., Colomba, L., and Garza, P. (2023). CaBuAr: California Burned Areas Dataset for Delineation. *IEEE Geosci. Remote Sens. Mag.*, 11(3):106–113. arXiv:2401.11519 [cs, eess].

[Chattopadhyay et al., 2018] Chattopadhyay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. (2018). Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, Lake Tahoe, NV, USA.

[Chen et al., 2022a] Chen, H., Dong, L., Yang, H., He, X., and Zhu, C. (2022a). Unsupervised Real-World Image Super-Resolution via Dual Synthetic-to-Realistic and Realistic-to-Synthetic Translations. *IEEE Signal Process. Lett.*, 29:1282–1286.

[Chen et al., 2024] Chen, H., He, Y., Zhang, L., Yang, W., Liu, Y., Gao, B., Zhang, Q., and Lu, J. (2024). A Multi-Input Channel U-Net Landslide Detection Method Fusing SAR Multisource Remote Sensing Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 17:1215–1232.

[Chen et al., 2017] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. arXiv:1706.05587 [cs].

[Chen et al., 2020a] Chen, R., Huang, W., Huang, B., Sun, F., and Fang, B. (2020a). Reusing Discriminators for Encoding: Towards Unsupervised Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8168–8177.

[Chen et al., 2020b] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020b). A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR. ISSN: 2640-3498.

[Chen and He, 2021] Chen, X. and He, K. (2021). Exploring Simple Siamese Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758.

[Chen and Bruzzone, 2022] Chen, Y. and Bruzzone, L. (2022). Self-Supervised Change Detection in Multiview Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–12.

[Chen and Sun, 2023] Chen, Z. and Sun, Q. (2023). Extracting Class Activation Maps From Non-Discriminative Features As Well. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3135–3144, Vancouver, Canada.

[Chen et al., 2022b] Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H., and Sun, Q. (2022b). Class Re-Activation Maps for Weakly-Supervised Semantic Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 959–968, New Orleans, LA, USA.

[Cheng et al., 2017] Cheng, Q., Shen, H., Zhang, L., and Peng, Z. (2017). Missing Information Reconstruction for Single Remote Sensing Images Using Structure-Preserving Global Optimization. *IEEE Signal Process. Lett.*, 24(8):1163–1167.

[Cho et al., 2014] Cho, K., Van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. arXiv: 1406.1078 [cs].

[Chollet, 2017] Chollet, F. (2017). Xception: Deep Learning With Depthwise Separable Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1251–1258, Honolulu, HI, USA.

[Cordts et al., 2016] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. (2016). The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, Las Vegas, NV, USA.

[Cruden et al., 1993] Cruden, D., Beltran, L., Krauter, E., Lefebvre, G., Ter-Stepanian, G., and Zhouyuan, Z. (1993). *Multilingual Landslide Glossary.* The International Geotechnical Societies UNESCO. Working Party for World Landslide Inventory. The Canadian Geotechnical Society. BiTech Publishers Ltd., Richmond.

[Dalsasso et al., 2021] Dalsasso, E., Denis, L., and Tupin, F. (2021). SAR2SAR: A Semi-Supervised Despeckling Algorithm for SAR Images. *IEEE J. Sel. Top. Appl. Earth Obs.*, 14:4321–4329.

[Darbaghshahi et al., 2022] Darbaghshahi, F. N., Mohammadi, M. R., and Soryani, M. (2022). Cloud Removal in Remote Sensing Images Using Generative Adversarial Networks and SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–9.

[de Vries et al., 2017] de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., and Courville, A. C. (2017). Modulating Early Visual Processing by Language. In *Advances in Neural Information Processing Systems*, volume 30, pages 1–11, Long Beach, CA, USA. Curran Associates, Inc.

[Delacourt et al., 2004] Delacourt, C., Allemand, P., Casson, B., and Vadon, H. (2004). Velocity Field of the "La Clapière" Landslide Measured by the Correlation of Aerial and QuickBird Satellite Images. *Geophys. Res. Lett.*, 31(15):1–5.

[Deline et al., 2022] Deline, P., Hewitt, K., Shugar, D., and Reznichenko, N. (2022). Chapter 9 - Rock Avalanches Onto Glaciers. In Davies, T., Rosser, N., and Shroder, J. F., editors, *Landslide Hazards, Risks, and Disasters (Second Edition)*, Hazards and Disasters Series, pages 269–333. Elsevier.

[Deng et al., 2009] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, Miami, FL, USA.

[Desai and Ramaswamy, 2020] Desai, S. and Ramaswamy, H. G. (2020). Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 983–991, Snowmass Village, CO, USA.

[Desrues et al., 2019] Desrues, M., Lacroix, P., and Brenguier, O. (2019). Satellite Pre-Failure Detection and In Situ Monitoring of the Landslide of the Tunnel du Chambon, French Alps. *Geosci.*, 9(7):313.

[Dhariwal and Nichol, 2021] Dhariwal, P. and Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc.

[Dosovitskiy et al., 2020] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[Du et al., 2022] Du, W.-L., Zhou, Y., Zhu, H., Zhao, J., Shao, Z., and Tian, X. (2022). A Semi-Supervised Image-to-Image Translation Framework for SAR–Optical Image Matching. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Dukai et al., 2019] Dukai, B., Ledoux, H., and Stoter, J. E. (2019). A Multi-Height Lod1 Model of All Buildings in the Netherlands. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-4-W8:51–57.

[Dumoulin et al., 2017] Dumoulin, V., Shlens, J., and Kudlur, M. (2017). A Learned Representation For Artistic Style. In *International Conference on Learning Representations*, pages 1–26, Toulon, France.

[Ebel et al., 2021] Ebel, P., Saha, S., and Zhu, X. X. (2021). Fusing Multi-Modal Data for Supervised Change Detection. *Int. Arch. Photogramm. Rem. Sens. Spatial Inf. Sci.*, XLIII-B3-2021:243–249.

[Erion et al., 2021] Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., and Lee, S.-I. (2021). Improving Performance of Deep Learning Models with Axiomatic Attribution Priors and Expected Gradients. *Nat. Mach. Intell.*, 3(7):620–631.

[Everingham et al., 2010] Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.*, 88(2):303–338.

[Fang et al., 2023] Fang, B., Chen, G., Kou, R., Paoletti, M. E., Haut, J. M., and Plaza, A. (2023). CIT: Content-Invariant Translation with Hybrid Attention Mechanism for Unsupervised Change Detection. *ISPRS J. Photogramm. Remote Sens.*, 204:321–339.

[Fang et al., 2020] Fang, Z., Wang, Y., Peng, L., and Hong, H. (2020). Integration of Convolutional Neural Network and Conventional Machine Learning Classifiers for Landslide Susceptibility Mapping. *Comput. Geosci.*, 139:104470.

[Farr et al., 2007] Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., and Alsdorf, D. (2007). The Shuttle Radar Topography Mission. *Rev. Geophys.*, 45(2):1–33.

[Fell et al., 2008] Fell, R., Corominas, J., Bonnard, C., Cascini, L., Leroi, E., and Savage, W. Z. (2008). Guidelines for Landslide Susceptibility, Hazard and Risk Zoning for Land Use Planning. *Eng. Geol.*, 102(3):85–98.

[Feng et al., 2022] Feng, J., Jiang, Q., Tseng, C.-H., Jin, X., Liu, L., Zhou, W., and Yao, S. (2022). A Deep Multitask Convolutional Neural Network for Remote Sensing Image Super-Resolution and Colorization. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15.

[Feranec et al., 2016] Feranec, J., Soukup, T., Hazeu, G., and Jaffrain, G. (2016). *European Landscape Dynamics: CORINE Land Cover Data*. CRC Press.

[Fruneau et al., 1996] Fruneau, B., Achache, J., and Delacourt, C. (1996). Observation and Modelling of the Saint-Étienne-De-Tinée Landslide Using SAR Interferometry. *Tectonophysics*, 265(3-4):181–190.

[Fu et al., 2019] Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., and Tao, D. (2019). Geometry-Consistent Generative Adversarial Networks for One-Sided Unsupervised Domain Mapping. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2422–2431, Long Beach, CA, USA. IEEE.

[Fu et al., 2023] Fu, J., Lu, T., Zhang, S., and Wang, G. (2023). UM-CAM: Uncertainty-weighted Multi-resolution Class Activation Maps for Weakly-supervised Fetal Brain Segmentation. In Greenspan, H., Madabhushi, A., Mousavi, P., Salcudean, S., Duncan, J., Syeda-Mahmood, T., and Taylor, R., editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 315–324, Vancouver, Canada.

[Fu et al., 2021] Fu, S., Xu, F., and Jin, Y.-Q. (2021). Reciprocal Translation Between SAR and Optical Remote Sensing Images with Cascaded-Residual Adversarial Networks. *Sci. China Inf. Sci.*, 64(2):1–15.

[Fuentes Reyes et al., 2019] Fuentes Reyes, M., Auer, S., Merkle, N., Henry, C., and Schmitt, M. (2019). SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks—Optimization, Opportunities and Limits. *Remote Sens.*, 11(17):2067.

[Ganin and Lempitsky, 2015] Ganin, Y. and Lempitsky, V. (2015). Unsupervised Domain Adaptation by Backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1180–1189, Lille, France. PMLR. ISSN: 1938-7228.

[Garnot et al., 2021] Garnot, V. S. F., Landrieu, L., and Chehata, N. (2021). Multi-Modal Temporal Attention Models for Crop Mapping from Satellite Time Series. *ISPRS J. Photogramm. Remote Sens.*, 187:294–305.

[Gastaldi, 2017] Gastaldi, X. (2017). Shake-Shake Regularization. arXiv:1705.07485 [cs].

[Ghorbanzadeh et al., 2019] Ghorbanzadeh, O., Blaschke, T., Gholamnia, K., Meena, S. R., Tiede, D., and Aryal, J. (2019). Evaluation of Different Machine Learning Methods and Deep-Learning Convolutional Neural Networks for Landslide Detection. *Remote Sens.*, 11(2):196.

[Ghorbanzadeh et al., 2021] Ghorbanzadeh, O., Crivellari, A., Ghamisi, P., Shahabi, H., and Blaschke, T. (2021). A Comprehensive Transferability Evaluation of U-Net and ResU-Net for Landslide Detection from Sentinel-2 Data (case Study Areas from Taiwan, China, and Japan). *Sci. Rep.*, 11(1):14629.

[Ghorbanzadeh et al., 2022a] Ghorbanzadeh, O., Xu, Y., Ghamisi, P., Kopp, M., and Kreil, D. (2022a). Landslide4Sense: Reference Benchmark Data and Deep Learning Models for Landslide Detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–17.

[Ghorbanzadeh et al., 2022b] Ghorbanzadeh, O., Xu, Y., Zhao, H., Wang, J., Zhong, Y., Zhao, D., Zang, Q., Wang, S., Zhang, F., Shi, Y., Zhu, X. X., Bai, L., Li, W., Peng, W., and Ghamisi, P. (2022b). The Outcome of the 2022 Landslide4Sense Competition: Advanced Landslide Detection from Multi-Source Satellite Imagery. *IEEE J. Sel. Top. Appl. Earth Obs.*, 15:9927–9942.

[Giardino et al., 2014] Giardino, R., Regmi, N., and Vitek, J. (2014). Rock Glacier. In *Encyclopedia of Ice Snow and Glaciers*. Springer, Dordrecht.

[Goldstein and Werner, 1998] Goldstein, R. M. and Werner, C. L. (1998). Radar Interferogram Filtering for Geophysical Applications. *Geophys. Res. Lett.*, 25(21):4035–4038.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. The MIT press, Cambridge, Mass.

[Goodfellow et al., 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27, pages 1–9, Montreal, QC, Canada. Curran Associates, Inc.

[Grill et al., 2020] Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., Piot, B., kavukcuoglu, k., Munos, R., and Valko, M. (2020). Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc.

[Gupta et al., 2019] Gupta, R., Goodman, B., Patel, N., Hosfelt, R., Sajeev, S., Heim, E., Doshi, J., Lucas, K., Choset, H., and Gaston, M. (2019). Creating xBD: A Dataset for Assessing Building Damage from Satellite Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 10–17, Long Beach, CA, USA.

[Guzzetti et al., 2020] Guzzetti, F., Gariano, S. L., Peruccacci, S., Brunetti, M. T., Marchesini, I., Rossi, M., and Melillo, M. (2020). Geographical Landslide Early Warning Systems. *Earth-Sci. Rev.*, 200:102973.

[He et al., 2016a] He, K., Zhang, X., Ren, S., and Sun, J. (2016a). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA.

[He and Yokoya, 2018] He, W. and Yokoya, N. (2018). Multi-Temporal Sentinel-1 and -2 Data Fusion for Optical Image Simulation. *ISPRS Int. J. Geo-Inf.*, 7(10):389.

[He et al., 2016b] He, X., Hong, Y., Vergara, H., Zhang, K., Kirstetter, P.-E., Gourley, J. J., Zhang, Y., Qiao, G., and Liu, C. (2016b). Development of a Coupled Hydrological-Geotechnical Framework for Rainfall-Induced Landslides Prediction. *J. Hydrol.*, 543:395–405.

[Hinton et al., 2012] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. arXiv:1207.0580 [cs].

[Hochreiter and Schmidhuber, 1997] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

[Hoffman et al., 2016] Hoffman, J., Gupta, S., and Darrell, T. (2016). Learning With Side Information Through Modality Hallucination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 826–834, Las Vegas, NV, USA.

[Hu et al., 2022] Hu, W.-S., Li, H.-C., Wang, R., Gao, F., Du, Q., and Plaza, A. (2022). Pseudo Complex-Valued Deformable ConvLSTM Neural Network with Mutual Attention Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.*, 60:1–17.

[Hungr et al., 2014] Hungr, O., Leroueil, S., and Picarelli, L. (2014). The Varnes Classification of Landslide Types, an Update. *Landslides*, 11(2):167–194.

[Ienco et al., 2017] Ienco, D., Gaetano, R., Dupaquier, C., and Maurel, P. (2017). Land Cover Classification via Multi-temporal Spatial Data by Recurrent Neural Networks. *IEEE Geosci. Remote Sensing Lett.*, 14(10):1685–1689.

[Ienco et al., 2019] Ienco, D., Interdonato, R., Gaetano, R., and Ho Tong Minh, D. (2019). Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for Land Cover Mapping Via a Multi-Source Deep Learning Architecture. *ISPRS J. Photogramm. Remote Sens.*, 158:11–22.

[Iizuka et al., 2016] Iizuka, S., Simo-Serra, E., and Ishikawa, H. (2016). Let There Be Color! Joint End-to-End Learning of Global and Local Image Priors for Automatic Image Colorization with Simultaneous Classification. *ACM Trans. Graph.*, 35(4):110:1–110:11.

[Imran et al., 2022] Imran, S., Tahir, M., Khalid, Z., and Uppal, M. (2022). A Deep Unfolded Prior-Aided RPCA Network for Cloud Removal. *IEEE Signal Process. Lett.*, 29:2048–2052.

[Inan and Rahman, 2023] Inan, M. S. K. and Rahman, I. (2023). Explainable AI Integrated Feature Selection for Landslide Susceptibility Mapping using TreeSHAP. *SN Comput. Sci.*, 4(5):482. arXiv:2201.03225 [cs].

[Interdonato et al., 2018] Interdonato, R., Ienco, D., Gaetano, R., and Ose, K. (2018). DUPLO: A DUal View Point Deep Learning Architecture for Time Series ClassificatiOn. *ISPRS J. Photogramm. Remote Sens.*, 149:91–104.

[Intrieri et al., 2019] Intrieri, E., Carla, T., Farina, P., Bardi, F., Ketizmen, H., and Casagli, N. (2019). Satellite Interferometry as a Tool for Early Warning and Aiding Decision Making in an Open-Pit Mine. *IEEE J. Sel. Top. Appl. Earth Obs.*, 12(12):5248–5258.

[Intrieri et al., 2013] Intrieri, E., Gigli, G., Casagli, N., and Nadim, F. (2013). Brief communication "Landslide Early Warning System: toolbox and general concepts". *Nat. Hazards Earth Syst. Sci.*, 13(1):85–90.

[Intrieri et al., 2017] Intrieri, E., Raspini, F., Fumagalli, A., Lu, P., Sara, I., Del Conte, S., Farina, P., Allievi, J., Ferretti, A., and Casagli, N. (2017). The Maoxian Landslide as Seen from Space: Detecting Precursors of Failure with Sentinel-1 Data. *Landslides*, 15:123–133.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 448–456, Lille, France. PMLR.

[Isola et al., 2017] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-To-Image Translation With Conditional Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, Honolulu, HI,USA.

[Jain et al., 2021] Jain, P., Schoen-Phelan, B., and Ross, R. (2021). Multi-Modal Self-Supervised Representation Learning for Earth Observation. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 3241–3244, Brussels, Belgium. ISSN: 2153-7003.

[Jiang et al., 2021] Jiang, P.-T., Zhang, C.-B., Hou, Q., Cheng, M.-M., and Wei, Y. (2021). LayerCAM: Exploring Hierarchical Class Activation Maps for Localization. *IEEE Trans. Image Process.*, 30:5875–5888.

[Jo and Yu, 2021] Jo, S. and Yu, I.-J. (2021). Puzzle-CAM: Improved Localization Via Matching Partial And Full Features. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 639–643, Anchorage, AK, USA.

[Jones et al., 2018] Jones, D. B., Harrison, S., Anderson, K., and Betts, R. A. (2018). Mountain Rock Glaciers Contain Globally Significant Water Stores. *Sci. Rep.*, 8(1):2834.

[Jones et al., 2019] Jones, D. B., Harrison, S., Anderson, K., and Whalley, W. B. (2019). Rock Glaciers and Mountain Hydrology: A Review. *Earth Sci. Res.*, 193:66–90.

[Kampffmeyer et al., 2018] Kampffmeyer, M., Salberg, A.-B., and Jenssen, R. (2018). Urban Land Cover Classification With Missing Data Modalities Using Deep Convolutional Neural Networks. *IEEE J. Sel. Top. Appl. Earth Obs.*, 11(6):1758–1768.

[Kandala et al., 2022] Kandala, H., Saha, S., Banerjee, B., and Zhu, X. X. (2022). Exploring Transformer and Multilabel Classification for Remote Sensing Image Captioning. *IEEE Geosci. Remote Sensing Lett.*, 19:1–5.

[Kaushik et al., 2021] Kaushik, S., Yan, Y., Ravanel, L., Magnin, F., and Trouvé, E. (2021). Visibility Analysis of Glaciers on Steep Slopes in the European ALPS Using Terrasar-X/PAZ Data. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5505–5508, Brussels, Belgium.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*, pages 1–15, San Diego, CA, USA. Version Number: 9.

[Kingma and Welling, 2014] Kingma, D. P. and Welling, M. (2014). Auto-Encoding Variational Bayes. arXiv: 1312.6114 [cs].

[Koeniguer et al., 2021] Koeniguer, E. C., Mayerowitz, A., Letheule, N., and Plyer, A. (2021). The Dream Database: A Multimode Database Including Optics, Radar, DSM (SRTM) and OSM Labels for Deep Machine Learning Purposes. In *2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS*, pages 5743–5746, Brussels, Belgium.

[Kofler et al., 2021] Kofler, C., Mair, V., Gruber, S., Todisco, M. C., Nettleton, I., Steger, S., Zebisch, M., Schneiderbauer, S., and Comiti, F. (2021). When Do Rock Glacier Fronts Fail? Insights from Two Case Studies in South Tyrol (italian Alps). *Earth Surf. Process. Landf.*, 46(7):1311–1327.

[Kraehenbuehl and Koltun, 2013] Kraehenbuehl, P. and Koltun, V. (2013). Parameter Learning and Convergent Inference for Dense Random Fields. In *Proceedings of the 30th International Conference on Machine Learning*, pages 513–521, Atlanta, GA, USA. PMLR.

[Krizhevsky et al., 2012] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 25, Lake Tahoe, NV, USA. Curran Associates, Inc.

[Krull et al., 2019] Krull, A., Buchholz, A.-O., and Jug, F. (2019). Noise2Void - Learning Denoising From Single Noisy Images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 2129–2137, Long Beach, CA, USA.

[Kumar et al., 2021] Kumar, S., Banerjee, B., and Chaudhuri, S. (2021). Improved Landcover Classification using Online Spectral Data Hallucination. *Neurocomputing*, 439:316–326.

[Kussul et al., 2017] Kussul, N., Lavreniuk, M., Skakun, S., and Shelestov, A. (2017). Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data. *IEEE Geosci. Remote Sens. Lett.*, 14(5):778–782. Conference Name: IEEE Geoscience and Remote Sensing Letters.

[Lacroix et al., 2018] Lacroix, P., Bièvre, G., Pathier, E., Kniess, U., and Jongmans, D. (2018). Use of Sentinel-2 Images for the Detection of Precursory Motions Before Landslide Failures. *Remote Sens. Environ.*, 215:507–516.

[Lacroix et al., 2022] Lacroix, P., Dini, B., and Cheaib, A. (2022). Measuring Kinematics of Slow-Moving Landslides from Satellite Images. In *Surface Displacement Measurement from Remote Sensing Images*, pages 315–338. John Wiley & Sons, Ltd.

[Lacroix et al., 2020] Lacroix, P., Handwerger, A. L., and Bièvre, G. (2020). Life and Death of Slow-Moving Landslides. *Nat. Rev. Earth Environ.*, 1(8):404–419.

[Lazecký et al., 2020] Lazecký, M., Spaans, K., González, P. J., Maghsoudi, Y., Morishita, Y., Albino, F., Elliott, J., Greenall, N., Hatton, E., Hooper, A., Juncu, D., McDougall, A., Walters, R. J., Watson, C. S., Weiss, J. R., and Wright, T. J. (2020). LiCSAR: An Automatic InSAR Tool for Measuring and Monitoring Tectonic and Volcanic Activity. *Remote Sensing*, 12(15):2430.

[Lea et al., 2016] Lea, C., Vidal, R., Reiter, A., and Hager, G. D. (2016). Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In Hua, G. and Jégou, H., editors, *Computer Vision – ECCV 2016 Workshops*, pages 47–54, Amsterdam, The Netherlands. Springer International Publishing.

[Lecun et al., 1998] Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE*, 86(11):2278–2324.

[Lee and Park, 2023] Lee, I. H. and Park, C. G. (2023). SAR-to-Virtual Optical Image Translation for Improving SAR Automatic Target Recognition. *IEEE Geosci. Remote Sens. Lett.*, 20:1–5.

[Lee et al., 2021] Lee, K. H., Park, C., Oh, J., and Kwak, N. (2021). LFI-CAM: Learning Feature Importance for Better Visual Explanation. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1335–1343, Montreal, QC, Canada.

[Lehtinen et al., 2018] Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., and Aila, T. (2018). Noise2Noise: Learning Image Restoration without Clean Data. In *International Conference on Machine Learning*, pages 2965–2974, Vienna, Austria. PMLR.

[Ley et al., 2018] Ley, A., Dhondt, O., Valade, S., Haensch, R., and Hellwich, O. (2018). Exploiting GAN-Based SAR to Optical Image Transcoding for Improved Classification via Deep Learning. In *EUSAR 2018; 12th European Conference on Synthetic Aperture Radar*, pages 1–6, Aachen, Germany.

[Li et al., 2021a] Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., and Chen, Q. (2021a). Involution: Inverting the Inherence of Convolution for Visual Recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12316–12325, Nashville, TN, USA. IEEE.

[Li et al., 2022a] Li, G., Pan, L., Qiu, L., Tan, Z., Xie, F., and Zhang, H. (2022a). A Two-Stage Shake-Shake Network for Long-Tailed Recognition of SAR Aerial View Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 249–256, New Orleans, LA, USA.

[Li et al., 2022b] Li, H., Gu, C., Wu, D., Cheng, G., Guo, L., and Liu, H. (2022b). Multiscale Generative Adversarial Network Based on Wavelet Feature Learning for SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–15.

[Li et al., 2022c] Li, H., Zhu, F., Zheng, X., Liu, M., and Chen, G. (2022c). MSCDUNet: A Deep Learning Framework for Built-Up Area Change Detection Integrating Multispectral, SAR, and VHR Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 15:5163–5176.

[Li et al., 2021b] Li, X., Du, Z., Huang, Y., and Tan, Z. (2021b). A Deep Translation (GAN) Based Change Detection Network for Optical and SAR Remote Sensing Images. *ISPRS J. Photogramm. Remote Sens.*, 179:14–34.

[Li et al., 2022d] Li, X., Lei, L., Sun, Y., and Kuang, G. (2022d). Dynamic-Hierarchical Attention Distillation With Synergetic Instance Selection for Land Cover Classification Using Missing Heterogeneity Images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16.

[Li et al., 2020] Li, Y., Fu, R., Meng, X., Jin, W., and Shao, F. (2020). A SAR-to-Optical Image Translation Method Based on Conditional Generation Adversarial Network (cGAN). *IEEE Access*, 8:60338–60343.

[Li and Hoiem, 2017] Li, Z. and Hoiem, D. (2017). Learning Without Forgetting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(12):2935–2947.

[Li et al., 2019] Li, Z., Yang, J., Liu, Z., Yang, X., Jeon, G., and Wu, W. (2019). Feedback Network for Image Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3867–3876, Long Beach, CA, USA.

[Liao et al., 2011] Liao, Z., Hong, Y., Kirschbaum, D., Adler, R. F., Gourley, J. J., and Wooten, R. (2011). Evaluation of TRIGRS (Transient Rainfall Infiltration and Grid-based Regional Slope-Stability Analysis)'s Predictive Skill for Hurricane-Triggered Landslides: A Case Study in Macon County, North Carolina. *Nat. Hazards*, 58:325–339.

[Lissak et al., 2020] Lissak, C., Bartsch, A., De Michele, M., Gomez, C., Maquaire, O., Raucoules, D., and Roulland, T. (2020). Remote Sensing for Assessing Landslides and Associated Hazards. *Surv. Geophys.*, 41:1391–1435.

[Liu and Lei, 2018] Liu, L. and Lei, B. (2018). Can SAR Images and Optical Images Transfer with Each Other? In *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pages 7019–7022, Valencia, Spain. ISSN: 2153-7003.

[Liu et al., 2021a] Liu, X., Hong, D., Chanussot, J., Zhao, B., and Ghamisi, P. (2021a). Modality Translation in Remote Sensing Time Series. *IEEE Trans. Geosci. Remote Sens.*, 60:1–14.

[Liu and Zhang, 2022] Liu, Y. and Zhang, L. (2022). Weakly Supervised Region of Interest Extraction Based on Uncertainty-Aware Self-Refinement Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16.

[Liu and Zhang, 2023] Liu, Y. and Zhang, L. (2023). Mutually Supervised Learning via Interactive Consistency for Geographic Object Segmentation from Weakly Labeled Remote Sensing Imagery. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2985–2989, Kuala Lumpur, Malaysia.

[Liu et al., 2021b] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021b). Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, Montreal, QC, Canada.

[Long et al., 2015] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, Boston, MA, USA.

[Luppino et al., 2024] Luppino, L. T., Hansen, M. A., Kampffmeyer, M., Bianchi, F. M., Moser, G., Jenssen, R., and Anfinsen, S. N. (2024). Code-Aligned Autoencoders for Unsupervised Change Detection in Multimodal Remote Sensing Images. *IEEE Trans. Neural Netw. Learn. Syst.*, 35(1):60–72.

[Luppino et al., 2021] Luppino, L. T., Kampffmeyer, M., Bianchi, F. M., Moser, G., Serpico, S. B., Jenssen, R., and Anfinsen, S. N. (2021). Deep Image Translation with an Affinity-Based Change Prior for Unsupervised Multimodal Change Detection. *IEEE Trans. Geosci. Remote Sensing*, 60:1–22.

[Ma et al., 2023] Ma, L., Wang, L., Zhao, C., E, J., and Ohtsuki, T. (2023). Multilayer Attention Mechanism for Change Detection in SAR Image Spatial-Frequency Domain. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2110–2114, Kuala Lumpur, Malaysia.

[Ma et al., 2020] Ma, P., Zhang, F., and Lin, H. (2020). Prediction of InSAR Time-Series Deformation Using Deep Convolutional Neural Networks. *Remote Sens. Lett.*, 11(2):137–145.

[Marcer et al., 2020] Marcer, M., Ringsø Nielsen, S., Ribeyre, C., Kummert, M., Duvillard, P.-A., Schoeneich, P., Bodin, X., and Genuite, K. (2020). Investigating the Slope Failures at the Lou Rock Glacier Front, French Alps. *Permafr. Periglac. Process.*, 31(1):15–30.

[Martin and Whalley, 1987] Martin, E. H. and Whalley, W. B. (1987). Rock Glaciers: Part 1: Rock Glacier Morphology: Classification and Distribution. *Prog. Phys. Geogr.: Earth Environ.*, 11(2):260–282. Publisher: SAGE Publications Ltd.

[McColl, 2022] McColl, S. T. (2022). Chapter 2 - Landslide causes and triggers. In Davies, T., Rosser, N., and Shroder, J. F., editors, *Landslide Hazards, Risks, and Disasters (Second Edition)*, Hazards and Disasters Series, pages 13–41. Elsevier.

[Meier et al., 1993] Meier, E., Frei, U., and Nüesch, D. (1993). Precise Terrain Corrected Geocoded Images. *SAR Geocoding: Data and Systems*, pages 173–185. Publisher: Karlsruhe, Germany: Herbert Wichmann Verlag GmbH.

[Merkle et al., 2018] Merkle, N., Auer, S., Muller, R., and Reinartz, P. (2018). Exploring the Potential of Conditional Adversarial Networks for Optical and SAR Image Matching. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 11(6):1811–1820.

[Merkle et al., 2017] Merkle, N., Fischer, P., Auer, S., and Müller, R. (2017). On the Possibility of Conditional Adversarial Networks for Multi-Sensor Image Matching. In *Int. Geosci. Remote Sens. Symp. (IGARSS)*, Int. Geosci. Remote Sens. Symp. (IGARSS), pages 2633–2636, Fort Worth, TX, USA.

[Mescheder et al., 2019] Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., and Geiger, A. (2019). Occupancy Networks: Learning 3D Reconstruction in Function Space. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4460–4470, Long Beach, CA, USA.

[Miao et al., 2023] Miao, W., Xu, Z., Geng, J., and Jiang, W. (2023). ECAE: Edge-Aware Class Activation Enhancement for Semisupervised Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.*, 61:1–14.

[Michel et al., 2014] Michel, G. P., Kobiyama, M., and Goerl, R. F. (2014). Comparative Analysis of SHALSTAB and SINMAP for Landslide Susceptibility Mapping in the Cunha River Basin, Southern Brazil. *J. Soils Sediments*, 14(7):1266–1277.

[Mirmazloumi et al., 2022] Mirmazloumi, S. M., Gambin, A. F., Wassie, Y., Barra, A., Palamà, R., Crosetto, M., Monserrat, O., and Crippa, B. (2022). InSAR Deformation Time Series Classification Using a Convolutional Neural Network. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 43:307–312.

[Mirza and Osindero, 2014] Mirza, M. and Osindero, S. (2014). Conditional Generative Adversarial Nets. arXiv: 1411.1784 [cs].

[Mondini et al., 2021] Mondini, A. C., Guzzetti, F., Chang, K.-T., Monserrat, O., Martha, T. R., and Manconi, A. (2021). Landslide Failures Detection and Mapping Using Synthetic Aperture Radar: Past, Present and Future. *Earth-Sci. Rev.*, 216:103574.

[Montanaro et al., 2022] Montanaro, A., Valsesia, D., Fracastoro, G., and Magli, E. (2022). Semi-Supervised Learning for Joint SAR and Multispectral Land Cover Classification. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Montello et al., 2022] Montello, F., Arnaudo, E., and Rossi, C. (2022). MMFlood: A Multimodal Dataset for Flood Delineation from Satellite Imagery. *IEEE Access*, 10:96774–96787.

[Morales et al., 2021] Morales, D., Talavera, E., and Remeseiro, B. (2021). Playing to Distraction: Towards a Robust Training of Cnn Classifiers Through Visual Explanation Techniques. *Neural Comput. Applic.*, 33(24):16937–16949.

[Mourchid et al., 2020] Mourchid, Y., Donias, M., and Berthoumieu, Y. (2020). Dual Color-Image Discriminators Adversarial Networks for Generating Artificial-SAR Colorized Images from Sentinel-1 Images. In *Machine Learning for Earth Observation Workshop (ECML/PKDD 2020)*, pages 1–10, Ghent, Belgium.

[Muhammad and Yeasin, 2020] Muhammad, M. and Yeasin, M. (2020). Eigen-CAM: Class Activation Map using Principal Components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, Glasgow, UK.

[Nagendra et al., 2022] Nagendra, S., Kifer, D., Mirus, B. B., Pei, T., Lawson, K., Li, W., Nguyen, H., Qiu, T., Tran, S., Shen, C., and Manjunatha, S. B. (2022). Constructing a Large-scale Landslide Database Across Heterogeneous Environments Using Task-Specific Model Updates. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 15:4349–4370.

[Naghibi et al., 2022] Naghibi, S. A., Khodaei, B., and Hashemi, H. (2022). An Integrated InSAR-Machine Learning Approach for Ground Deformation Rate Modeling in Arid Areas. *J. Hydrol.*, 608:127627.

[Naidu et al., 2020] Naidu, R., Ghosh, A., Maurya, Y., K, S. R. N., and Kundu, S. S. (2020). IS-CAM: Integrated Score-CAM for Axiomatic-Based Explanations. arXiv:2010.03023 [cs].

[Nair and Hinton, 2010] Nair, V. and Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 807–814, Madison, WI, USA. Omnipress.

[Nava et al., 2021] Nava, L., Monserrat, O., and Catani, F. (2021). Improving Landslide Detection on SAR Data through Deep Learning. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Ndikumana et al., 2018] Ndikumana, E., Ho Tong Minh, D., Baghdadi, N., Courault, D., and Hossard, L. (2018). Deep Recurrent Neural Network for Agricultural Classification using multitemporal SAR Sentinel-1 for Camargue, France. *Remote Sens.*, 10(8):1217.

[Nielsen, 2015] Nielsen, M. (2015). *Neural Networks and Deep Learning*. Determination Press.

[Odena et al., 2017] Odena, A., Olah, C., and Shlens, J. (2017). Conditional Image Synthesis with Auxiliary Classifier GANs. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2642–2651, Sydney, Australia. PMLR.

[Orlandi et al., 2022] Orlandi, D., Galatolo, F. A., Cimino, M. G. C. A., Rosa, A. L., Pagli, C., and Perilli, N. (2022). Enhancing Land Subsidence Awareness via InSAR Data and Deep Transformers. In *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 98–103, Salerno, Italy. ISSN: 2379-1675.

[Pan et al., 2023] Pan, Y., Ahmed Khan, I., and Meng, H. (2023). SAR-to-Optical Image Translation Using Multi-Stream Deep ResCNN of Information Reconstruction. *Expert Syst. Appl.*, 224:120040.

[Piva and Dubbelman, 2023] Piva, F. J. and Dubbelman, G. (2023). Exploiting Image Translations Via Ensemble Self-Supervised Learning for Unsupervised Domain Adaptation. *Comput. Vis. Image Underst.*, 234:103745.

[Prada-Sarmiento et al., 2019] Prada-Sarmiento, L. F., Cabrera, M. A., Camacho, R., Estrada, N., and Ramos-Cañón, A. M. (2019). The Mocoa Event on March 31 (2017): Analysis of a Series of Mass Movements in a Tropical Environment of the Andean-Amazonian Piedmont. *Landslides*, 16(12):2459–2468.

[Qiao et al., 2023] Qiao, W., Shen, L., Wang, J., Yang, X., and Li, Z. (2023). A Weakly Supervised Semantic Segmentation Approach for Damaged Building Extraction From Postearthquake High-Resolution Remote-Sensing Images. *IEEE Geosci. Remote Sens. Lett.*, 20:1–5.

[Qin et al., 2022] Qin, J., Wu, J., Xiao, X., Li, L., and Wang, X. (2022). Activation Modulation and Recalibration Scheme for Weakly Supervised Semantic Segmentation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 2117–2125. Number: 2.

[Raissi et al., 2019] Raissi, M., Perdikaris, P., and Karniadakis, G. E. (2019). Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations. *J. Comput. Phys.*, 378:686–707.

[Rambour et al., 2020] Rambour, C., Audebert, N., Koeniguer, E., Le Saux, B., Crucianu, M., and Datcu, M. (2020). Flood Detection in Time Series of Optical and SAR Images. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, 43(B2):1343–1346.

[Ray et al., 2020] Ray, R. L., Lazzari, M., and Olutimehin, T. (2020). *Remote Sensing Approaches and Related Techniques to Map and Study Landslides*, volume 2 of *Landslides-Investig. Monit.* IntechOpen.

[Rebuffi et al., 2020] Rebuffi, S.-A., Fong, R., Ji, X., and Vedaldi, A. (2020). There and Back Again: Revisiting Backpropagation Saliency Methods. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8836–8845, Seattle, WA, USA.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science, pages 234–241, Munich, Germany. Springer International Publishing.

[Rosenblatt, 1962] Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books.

[Ross et al., 2017] Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2662–2670, Melbourne, Australia.

[Rouyet et al., 2021] Rouyet, L., Lilleøren, K. S., Böhme, M., Vick, L. M., Delaloye, R., Etzelmüller, B., Lauknes, T. R., Larsen, Y., and Blikra, L. H. (2021). Regional Morpho-Kinematic Inventory of Slope Movements in Northern Norway. *Front. Earth Sci.*, 9:681088.

[Rumelhart et al., 1986] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *Nature*, 323(6088):533–536. Publisher: Nature Publishing Group.

[Scheibenreif et al., 2022] Scheibenreif, L., Hanna, J., Mommert, M., and Borth, D. (2022). Self-supervised Vision Transformers for Land-cover Segmentation and Classification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1421–1430, New Orleans, LA, USA. IEEE.

[Schmitt et al., 2018] Schmitt, M., Hughes, L. H., and Zhu, X. X. (2018). The SEN1-2 Dataset for Deep Learning in Sar-Optical Data Fusion. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, IV-1:141–146.

[Schmitt et al., 2019] Schmitt, M., Hughes, L. M., Qiu, C., and Zhu, X. X. (2019). SEN12MS – A Curated Dataset of Georeferenced Multi-Spectral Sentinel-1/2 Imagery for Deep Learning and Data Fusion. arXiv:1906.07789 [cs].

[Schöttl, 2020] Schöttl, A. (2020). A Light-weight Method to Foster the (Grad)CAM Interpretability and Explainability of Classification Networks. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)*, pages 348–351.

[Selvaraju et al., 2017] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, Honolulu, HI, USA.

[Shermeyer et al., 2020] Shermeyer, J., Hogan, D., Brown, J., Van Etten, A., Weir, N., Pacifici, F., Hansch, R., Bastidas, A., Soenen, S., Bacastow, T., and Lewis, R. (2020). SpaceNet 6: Multi-Sensor All Weather Mapping Dataset. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA.

[Shi et al., 2016] Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, Las Vegas, NV, USA.

[Shi et al., 2015] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. (2015). Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In *Advances in Neural Information Processing Systems*, volume 28, Montreal, QC, Canada. Curran Associates, Inc.

[Shimosato and Ukita, 2021] Shimosato, K. and Ukita, N. (2021). Multi-Modal Data Fusion for Land-Subsidence Image Improvement in PSInSAR Analysis. *IEEE Access*, 9:141970–141980.

[Simonyan et al., 2014] Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034 [cs].

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs].

[Sohl-Dickstein et al., 2015] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265, Lille, France. PMLR.

[Sulla-Menashe et al., 2019] Sulla-Menashe, D., Gray, J. M., Abercrombie, S. P., and Friedl, M. A. (2019). Hierarchical mapping of annual global land cover 2001 to present: The MODIS Collection 6 Land Cover product. *Remote Sens. Environ.*, 222:183–194.

[Sumbul et al., 2019] Sumbul, G., Charfuelan, M., Demir, B., and Markl, V. (2019). Bigearthnet: A Large-Scale Benchmark Archive for Remote Sensing Image Understanding. In *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901–5904, Yokohama, Japan. ISSN: 2153-7003.

[Sumbul et al., 2021] Sumbul, G., de Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., and Markl, V. (2021). BigEarthNet-MM: A Large-Scale, Multimodal, Multilabel Benchmark Archive for Remote Sensing Image Classification and Retrieval [Software and Data Sets]. *IEEE Geosci. Remote Sens. Mag.*, 9(3):174–180.

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, Boston, MA, USA.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, Las Vegas, NV, USA.

[Thollard et al., 2021] Thollard, F., Clesse, D., Doin, M.-P., Donadieu, J., Durand, P., Grandin, R., Lasserre, C., Laurent, C., Deschamps-Ostanciaux, E., Pathier, E., Pointal, E., Proy, C., and Specht, B. (2021). FLATSIM: The ForM@Ter LArge-Scale Multi-Temporal Sentinel-1 InterferoMetry Service. *Remote Sens.*, 13(18):3734.

[Tran et al., 2022] Tran, Q.-A., Rogstad, A., Nordal, G. G. S., and Alene, G. H. (2022). Preliminary Study of 3D Large Deformation Modeling of the Gjerdrum 2020 Sensitive Clay Landslide. In *Proceedings in the 8th Canadian Conference on Geotechnique and Natural Hazards*, pages 1–8, Quebec, Canada.

[Ullo et al., 2019] Ullo, S. L., Langenkamp, M. S., Oikarinen, T. P., Del Rosso, M. P., Sebastianelli, A., Piccirillo, F., and Sica, S. (2019). Landslide Geohazard Assessment With Convolutional Neural Networks Using Sentinel-2 Imagery Data. In *2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 9646–9649, Yokohama, Japan.

[Valade et al., 2019] Valade, S., Ley, A., Massimetti, F., D'Hondt, O., Laiolo, M., Coppola, D., Loibl, D., Hellwich, O., and Walter, T. R. (2019). Towards Global Volcano Monitoring Using Multisensor Sentinel Missions and Artificial Intelligence: The MOUNTS Monitoring System. *Remote Sens.*, 11(13):1528.

[Varnes, 1978] Varnes, D. J. (1978). Slope Movement Types and Processes. *Transportation Research Board Special Report*, 176:11–33.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. U., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30, Long Beach, CA, USA. Curran Associates, Inc.

[Wang et al., 2020a] Wang, H., Wang, Z., Du, M., Yang, F., Zhang, Z., Ding, S., Mardziel, P., and Hu, X. (2020a). Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, Seattle, WA, USA.

[Wang et al., 2022a] Wang, H., Zhang, Z., Hu, Z., and Dong, Q. (2022a). SAR-to-Optical Image Translation With Hierarchical Latent Features. *IEEE Trans. Geosci. Remote Sens.*, 60:1–12.

[Wang et al., 2022b] Wang, L., Wang, J., Liu, Z., Zhu, J., and Qin, F. (2022b). Evaluation of a Deep-Learning Model for Multispectral Remote Sensing of Land Use and Crop Classification. *Crop J.*, 10(5):1435–1451.

[Wang et al., 2019] Wang, Q., Yuan, Z., Du, Q., and Li, X. (2019). GETNET: A General End-to-End 2-D CNN Framework for Hyperspectral Image Change Detection. *IEEE Trans. Geosci. Remote Sensing*, 57(1):3–13.

[Wang et al., 2022c] Wang, X., Yi, J., Guo, J., Song, Y., Lyu, J., Xu, J., Yan, W., Zhao, J., Cai, Q., and Min, H. (2022c). A Review of Image Super-Resolution Approaches Based on Deep Learning and Applications in Remote Sensing. *Remote Sensing*, 14(21):5423.

[Wang et al., 2023] Wang, Y., Deng, S., Yuan, K., Schaefer, G., Liu, X., and Fang, H. (2023). A Novel Class Activation Map for Visual Explanations in Multi-Object Scenes. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 2615–2619, Kuala Lumpur, Malaysia.

[Wang et al., 2020b] Wang, Y., Zhang, J., Kan, M., Shan, S., and Chen, X. (2020b). Self-Supervised Equivariant Attention Mechanism for Weakly Supervised Semantic Segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12272–12281, Seattle, WA, USA.

[Wang and Bunyak, 2023] Wang, Y. Y. and Bunyak, F. (2023). DFT-CAM: Discrete Fourier Transform Driven Class Activation Map. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 500–504, Kuala Lumpur, Malaysia.

[Wang et al., 2022d] Wang, Z., Ma, Y., and Zhang, Y. (2022d). Hybrid cGAN: Coupling Global and Local Features for SAR-to-Optical Image Translation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–16.

[Wargnier-Dauchelle et al., 2023] Wargnier-Dauchelle, V., Grenier, T., Durand-Dubief, F., Cotton, F., and Sdika, M. (2023). A Weakly Supervised Gradient Attribution Constraint for Interpretable Classification and Anomaly Detection. *IEEE Trans. Med. Imaging*, 42(11):3336–3347.

[Wasowski and Bovenga, 2022] Wasowski, J. and Bovenga, F. (2022). Chapter 11 - Remote sensing of landslide motion with emphasis on satellite multi-temporal interferometry applications: an overview. In Davies, T., Rosser, N., and Shroder, J. F., editors, *Landslide Hazards, Risks, and Disasters (Second Edition)*, Hazards and Disasters Series, pages 365–438. Elsevier.

[Wei et al., 2023a] Wei, J., Zou, H., Sun, L., Cao, X., He, S., Liu, S., and Zhang, Y. (2023a). CFRWD-GAN for SAR-to-Optical Image Translation. *Remote Sens.*, 15(10):2547.

[Wei et al., 2023b] Wei, S., Luo, Y., Ma, X., Ren, P., and Luo, C. (2023b). MSH-Net: Modality-Shared Hallucination With Joint Adaptation Distillation for Remote Sensing Image Classification Using Missing Modalities. *IEEE Trans. Geosci. Remote Sens.*, 61:1–15.

[Wei et al., 2021] Wei, X., Zhang, L., Luo, J., and Liu, D. (2021). A Hybrid Framework Integrating Physical Model and Convolutional Neural Network for Regional Landslide Susceptibility Mapping. *Nat. Hazards*, 109:471–497.

[Woo et al., 2018] Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 3–19, Munich, Germany. Springer International Publishing.

[Xia et al., 2024] Xia, Y., He, W., Huang, Q., Chen, H., Huang, H., and Zhang, H. (2024). SOSSF: Landsat-8 Image Synthesis on the Blending of Sentinel-1 and MODIS Data. *IEEE Trans. Geosci. Remote Sens.*, 62:1–19.

[Xu et al., 2024] Xu, Y., Ouyang, C., Xu, Q., Wang, D., Zhao, B., and Luo, Y. (2024). CAS Landslide Dataset: A Large-Scale and Multisensor Dataset for Deep Learning-Based Landslide Detection. *Sci. Data*, 11(1):12.

[Yang et al., 2022a] Yang, R., Zhang, F., Xia, J., and Wu, C. (2022a). Landslide Extraction Using Mask R-CNN with Background-Enhancement Method. *Remote Sens.*, 14(9):2206.

[Yang et al., 2022b] Yang, X., Zhao, J., Wei, Z., Wang, N., and Gao, X. (2022b). SAR-to-Optical Image Translation Based on Improved cGAN. *Pattern Recognit.*, 121:108208.

[Yun et al., 2019] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, Seoul, Korea (South).

[Zhang et al., 2020a] Zhang, B., Xiao, J., Wei, Y., Sun, M., and Huang, K. (2020a). Reliability Does Matter: An End-to-End Weakly Supervised Semantic Segmentation Approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12765–12772.

[Zhang et al., 2018a] Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. (2018a). mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*, Vancouver, Canada.

[Zhang et al., 2023a] Zhang, H., Torres, F., Sicre, R., Avrithis, Y., and Ayache, S. (2023a). Opti-CAM: Optimizing saliency maps for interpretability. arXiv:2301.07002 [cs].

[Zhang et al., 2024] Zhang, H., Yang, Y.-J., and Zeng, W. (2024). Self-Supervised Multi-Scale Semantic Consistency Regularization for Unsupervised Image-to-Image Translation. *Comput. Vis. Image Underst.*, 241:103950.

[Zhang et al., 2023b] Zhang, J., Liu, Z., Jiang, W., Liu, Y., Zhou, X., and Li, X. (2023b). Application of Deep Generative Networks for SAR/ISAR: A Review. *Artif. Intell. Rev.*, 56(10):11905–11983.

[Zhang et al., 2020b] Zhang, J., Zhou, J., and Lu, X. (2020b). Feature-Guided SAR-to-Optical Image Translation. *IEEE Access*, 8:70925–70937.

[Zhang and Li, 2021] Zhang, K. and Li, Y. (2021). Single Image Dehazing via Semi-Supervised Domain Translation and Architecture Search. *IEEE Signal Process. Lett.*, 28:2127–2131.

[Zhang and Ma, 2021] Zhang, L. and Ma, J. (2021). Salient Object Detection Based on Progressively Supervised Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.*, 59(11):9682–9696.

[Zhang et al., 2022] Zhang, T., Gao, F., Dong, J., and Du, Q. (2022). Remote Sensing Image Translation via Style-Based Recalibration Module and Improved Style Discriminator. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Zhang and Xu, 2019] Zhang, W. and Xu, M. (2019). Translate SAR Data into Optical Image Using IHS and Wavelet Transform Integrated Fusion. *J. Indian Soc. Remote Sens.*, 47(1):125–137.

[Zhang et al., 2018b] Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. (2018b). Adversarial Complementary Learning for Weakly Supervised Object Localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, Salt Lake City, UT, USA.

[Zhang et al., 2018c] Zhang, Z., Vosselman, G., Gerke, M., Tuia, D., and Yang, M. Y. (2018c). Change Detection between Multimodal Remote Sensing Data Using Siamese CNN. *arXiv:1807.09562 [cs]*. arXiv: 1807.09562.

[Zhang et al., 2017] Zhang, Z., Wang, H., Xu, F., and Jin, Y.-Q. (2017). Complex-Valued Convolutional Neural Network and Its Application in Polarimetric SAR Image Classification. *IEEE Trans. Geosci. Remote Sens.*, 55(12):7177–7188.

[Zhao and Lu, 2018] Zhao, C. and Lu, Z. (2018). Remote Sensing of Landslides—A Review. *Remote Sens.*, 10(2):279.

[Zhao and Wang, 2023] Zhao, G. and Wang, W. (2023). Change Detection for Remote Sensing Images Based on Semantic Prototypes and Contrastive Learning. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 865–869, Kuala Lumpur, Malaysia.

[Zhao et al., 2023] Zhao, T., Wang, L., Zhao, C., Liu, T., and Ohtsuki, T. (2023). Heterogeneous Image Change Detection Based on Deep Image Translation and Feature Refinement-Aggregation. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1705–1709, Kuala Lumpur, Malaysia.

[Zhao et al., 2022a] Zhao, Y., Celik, T., Liu, N., and Li, H.-C. (2022a). A Comparative Analysis of GAN-Based Methods for SAR-to-Optical Image Translation. *IEEE Geosci. Remote Sens. Lett.*, 19:1–5.

[Zhao et al., 2022b] Zhao, Y., Feng, G., Wang, Y., Wang, X., Wang, Y., Lu, H., Xu, W., and Wang, H. (2022b). A New Algorithm for Intelligent Detection of Geohazards Incorporating Attention Mechanism. *Int. J. Appl. Earth Obs. Geoinf.*, 113:102988.

[Zheng et al., 2021] Zheng, Z., Ma, A., Zhang, L., and Zhong, Y. (2021). Deep Multisensor Learning for Missing-Modality All-Weather Mapping. *ISPRS J. Photogramm. Remote Sens.*, 174:254–264.

[Zhou et al., 2016] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning Deep Features for Discriminative Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, Las Vegas, NV, USA.

[Zhou et al., 2018] Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., and Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. In Stoyanov, D., Taylor, Z., Carneiro, G., Syeda-Mahmood, T., Martel, A., Maier-Hein, L., Tavares, J. M. R. S., Bradley, A., Papa, J. P., Belagiannis, V., Nascimento, J. C., Lu, Z., Conjeti, S., Moradi, M., Greenspan, H., and Madabhushi, A., editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Granada, Spain. Springer International Publishing.

[Zhu et al., 2017] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision*, Proc. IEEE Int. Conf. Comput. Vis., pages 2223–2232, Venice, Italy.

[Zhu et al., 2023] Zhu, L., She, Q., Chen, Q., Meng, X., Geng, M., Jin, L., Zhang, Y., Ren, Q., and Lu, Y. (2023). Background-Aware Classification Activation Map for Weakly Supervised Object Localization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(12):14175–14191.

Résumé Long en Français

Introduction

La thèse présentée dans ce manuscrit vise à concilier trois mondes de la recherche : les géosciences, l'imagerie satellitaire ainsi que l'apprentissage automatique. En effet, les travaux menés ici visent à :

"Simplement détecter des glissements de terrain et phénomènes semblables sur des images satellitaires à l'aide d'intelligence artificielle."

Et même si cette formulation vulgarisée peut sembler "simple", elle cache de nombreux défis liés à la fois aux données utilisées, aux phénomènes recherchés ainsi qu'aux méthodes à implémenter. De fait, dans cette thèse l'objectif principal est d'assurer la cohabitation de chaque thématique et de trouver des solutions pour que les défauts de l'une soient compensés par les atouts des autres. Mais afin de compenser de tels défauts, encore faut il les connaître et les comprendre.

Instabilités gravitaires

Le terme *instabilité gravitaire* employé dans ces travaux réfère à l'ensemble des mouvements de terrain induits par une force gravitaire. Les auteurs de [Varnes, 1978] ont rapidement établi une catégorisation de ces phénomènes qui a été reprise et mise à jour dans [Hungr et al., 2014]. Parmi l'ensemble de ces instabilités, ces travaux se concentrent sur les phénomènes liés à la pente sous-jacente et en particulier aux glissements de terrain et aux glaciers rocheux.

Les glissements de terrain sont considérés comme des masses de terre en mouvement le long d'une surface de rupture. Néanmoins, ils ne sont composés que de matériaux solides et sont rarement à l'origine de leur propre mouvement, excepté sur des pentes particulièrement fortes. Malgré tout, des évènements déclencheurs peuvent être à l'origine de leur mise en mouvement ou même pire, de leur déclenchement. En effet, suite à des épisodes de pluie intense, l'eau peut s'infiltrer dans les couches souterraines, les fluidifier et entraîner l'ensemble des couches supérieures à suivre le flux d'eau et donc à se déplacer en bas de la pente. De même des évènements tels que des séismes ou des exploitations minières peuvent fragiliser la structure des couches souterraines. Celles-ci sont donc susceptibles de se briser et de déclencher un glissement entraînant de nouveau les couches supérieures en bas de la pente.

Si les glaciers rocheux sont aussi sujets aux évènements précédents, ils sont en plus par nature beaucoup plus susceptibles d'être mis en mouvement. En effet, la présence d'un ancien glacier sous les blocs rocheux ou celle de glace dans les interstices rocheux comme indiqué sur la Figure 2.4 les rendent particulièrement sujets aux déplacements lorsque la glace se met à fondre. Et même si ces instabilités n'étaient étudiées que comme des sources d'eau naturelles dans un premier temps [Martin and Whalley, 1987, Jones et al., 2019], depuis quelques années elles sont considérées comme de potentiels dangers suite à une série d'effondrement [Bodin et al., 2017, Kofler et al., 2021, Marcer et al., 2020].

De fait le suivi de ces phénomènes est primordial pour éviter qu'ils ne deviennent des catastrophes naturelles détruisant les infrastructures et menaçant des vies. Dans cette thèse, deux stades du suivi d'instabilités sont considérés : leur détection pre- et post-déclenchement. Dans le premier cas, l'objectif est d'identifier des phénomènes dont la vitesse peut varier de quelques centimètres à quelques mètres par an [Lacroix et al., 2020]. Cet inventaire de potentiels dangers peut ensuite être utilisé par les autorités locales

afin de faire une cartographie du risque et ainsi prendre des mesures pour limiter les potentiels dégâts que causerait le déclenchement de ces instabilités. Dans le second cas, l'objectif est d'identifier les instabilités suite à leur déclenchement. À ce stade du processus, la prévention n'est plus de rigueur : l'action prend le dessus. Néanmoins, pour assurer l'envoi et la sécurité des équipes de secours sur place, il est primordial de bien connaître l'étendue des dégâts causés par ces glissements.

Un grand défi lié à ces phénomènes gravitaires réside dans leur variabilité. En effet, de part la complexité de la physique sous-jacente, il n'existe pas de modèle global permettant de prédire le déclenchement, l'accélération ou même la mise en mouvement d'une instabilité. Si plusieurs modèles physiques se basent largement sur des comportements de particules au sein d'une pente "infinie" conditionnés par la quantité et le flux de l'eau [Liao et al., 2011, Michel et al., 2014, He et al., 2016b, An et al., 2016], ils ne permettent pas de refléter toute la complexité des phénomènes sous-jacents. De fait, les études portant sur les instabilités gravitaires se sont lancées dans une véritable quête à la compréhension et aux indices interprétables pour mieux identifier le comportement de ces phénomènes.

Images de télédétection

Au delà des études géophysiques individuelles des instabilités gravitaires, cette thèse propose une étude plus globale afin de les détecter à large échelle. Pour ce faire, l'attention s'est portée sur des considérations géomorphologiques de surface et visibles à partir d'imagerie de télédétection, en particulier satellitaire. En effet, les images de télédétection ont la faculté de couvrir de très vastes régions. En particulier, les images obtenues par les satellites des constellations Sentinel-1 et Sentinel-2 utilisées dans cette thèse couvrent respectivement prêt de 49.000 et 12.000 km^2 par image téléchargée depuis le site de l'Agence Spatiale Européenne (ESA en anglais).

Ces deux constellations sont utilisées pour deux raisons principales : la fréquence des acquisitions ainsi que l'accès libre à la bibliothèque d'images. En effet, les contributions proposées dans cette thèse, à savoir la détection de changements ainsi que le suivi de mouvement, sont essentiellement d'ordre bitemporel, *i.e.* nécessitant l'acquisition de deux images pour être effectives. De fait, la fréquence des acquisitions des constellations Sentinel-1 (6 jours) et Sentinel-2 (5 jours) les rendent particulièrement efficaces pour une détection précoce d'instabilités ou pour une réaction rapide suite à leur déclenchement. Du point de vue de l'accessibilité, les travaux menés dans cette thèse visant à être exploitables et généralisables dans d'autres contextes et régions, il semblait donc nécessaire d'utiliser des données accessibles à tous.

Mais si ces deux constellations sont utilisées conjointement c'est surtout parce qu'elles sont complémentaires pour la compréhension de la couverture du sol. La constellation Sentinel-2 est une constellation de deux satellites à capteur multispectral. Cette caractéristique lui permet d'acquérir des images du sol dans différentes longueurs d'onde : à la fois dans le domaine du visible (Bleu, Vert et Rouge) mais aussi dans plusieurs bandes infra-rouges. Ceci rend donc les images particulièrement instructives et compréhensibles pour un utilisateur final, mais aussi bien plus facilement exploitables pour des méthodes de détection de changements [Nava et al., 2021] contrairement aux images radar acquises par Sentinel-1. En effet, cette constellation était composée avant décembre 2021 de deux satellites Radar à Synthèse d'Ouverture (SAR en anglais). Ces images SAR sont acquises suite à l'émission par le satellite d'une onde électromagnétique (en bande C pour Sentinel-1) et à l'acquisition du signal rétrodiffusé par le sol en direction du satellite. De fait, très peu d'informations radiométriques sont perceptibles sur les images SAR, ce qui rend leur interprétation beaucoup plus compliquée pour des non-experts. Mais malgré tout, l'information délivrée par ces images permet de bien mieux cerner la structuralité de la région imagée. De plus, une caractéristique inhérente à l'imagerie SAR est d'être "tout temps", donc acquises de jour comme de nuit et indépendamment de la couverture nuageuse.

Les instabilités gravitaires étant majoritairement déclenchées par des intempéries comme le montre le rapport du World Bank Group[4], ceci implique la présence d'une forte couverture nuageuse, privant de l'utilisation des images multispectrales pour une interprétation rapide des dégâts. Au contraire, les images SAR étant acquises régulièrement, il est possible de les exploiter dès leur acquisition pour permettre une

[4]https://datacatalogfiles.worldbank.org/ddh-published/0037584/DR0045411/global-landslide-hazard-map-report.pdf?versionId=2023-01-18T20:43:00.6156776Z, visité le 26 juin 2024.

réaction rapide à de potentielles menaces déclenchées par ces fortes pluies. Malgré tout, une autre caractéristique des instabilités gravitaires est leur localisation : généralement au sein de régions montagneuses avec de fortes pentes. Ce détail a pour conséquence de limiter également l'utilisation des images SAR. En effet, de part le processus d'acquisition en distance par rapport à la Ligne de Visée (LOS en anglais) du satellite, les acquisitions SAR sont fortement perturbées par la topographie du terrain. En particulier, des distorsions géométriques peuvent apparaître telles que les chevauchements ou ombres radar [Kaushik et al., 2021], empêchant l'interprétation des zones sujettes à ces distorsions.

Enfin, il est également important de tenir compte de la variabilité des déplacements des instabilités gravitaires lentes. En effet, des déplacements annuels aussi bien centimétriques que métriques ne peuvent décemment être visibles sur des images satellitaires telles que Sentinel-1 et Sentinel-2 de part leur résolution décamétrique. Néanmoins, les images SAR peuvent être exploitées non seulement en amplitude mais aussi en phase : du fait du processus d'émission et réception d'une même onde, chaque pixel SAR a en réalité une valeur complexe. Or, le déphasage étant lié au temps de parcourt aller-retour de l'onde, un déphasage suite à un mouvement de terrain entre deux acquisitions doit permettre de le détecter. De plus, la longueur d'onde utilisée pour Sentinel-1 étant de 5.55 cm, il est possible de détecter des mouvements jusqu'à 2.27 cm le long de la LOS dans le cas où un déphasage de 2π est observé. Cette simple observation a mené à la génération d'interférogrammes permettant de calculer une carte de différence de phase et de cohérence révélatrice de la fiabilité du mouvement détecté.

Ainsi donc, la couverture nuageuse, les distorsions géométriques, la résolution décamétrique ainsi que l'interprétabilité des images sont autant de défis à prendre en compte pour pouvoir exploiter au mieux les données satellitaires à des fins de détection d'instabilités gravitaires.

Apprentissage automatique

L'imagerie de télédétection a déjà été largement utilisée pour l'étude de phénomènes gravitaires. En particulier certains auteurs se sont concentrés sur la détection des mouvements lents par corrélation d'images optiques [Delacourt et al., 2004, Desrues et al., 2019] et à l'aide d'interférogrammes [Bertone et al., 2022, Burrows et al., 2019], quand d'autres visent à établir des inventaires d'instabilités gravitaires aussi bien optiques [Jones et al., 2018, Lissak et al., 2020] que radar [Bodin, 2020].

Avec le développement très rapide et les gains en performances de l'intelligence artificielle et en particulier de l'apprentissage profond en vision par ordinateur, ces méthodes se sont donc naturellement étendues à l'imagerie de télédétection et aux instabilités gravitaires. Mais la combinaison de ces trois thématiques est plus rarement étudiée pour plusieurs raisons inhérentes aux approches d'apprentissage profond.

Premièrement, les approches dites neuronales requièrent un très grand nombre de données pour pouvoir atteindre des performances optimales. Si la quantité d'images satellitaires correspond parfaitement aux besoins de part la variété des satellites, des modalités et des acquisitions, les données relatives aux instabilités gravitaires sont beaucoup plus limitées. En effet, les approches d'apprentissage automatique sont majoritairement guidées par des données annotées, *i.e.* dont la tâche cible est connue, pour mener un entraînement supervisé. Ceci requiert donc la mise en place de jeux de données annotés par des experts dans le domaine pour pouvoir guider au mieux les réseaux neuronaux à accomplir cette tâche. Malheureusement, mis à part le jeu de données Landslide4Sense [Ghorbanzadeh et al., 2022a] proposé sous la forme d'un défi à la communauté pour l'inventaire monotemporel optique de glissements de terrain, peu de jeux de données ont été créés pour étudier les instabilités gravitaires, ce qui limite nécessairement l'utilisation de l'apprentissage profond pour des détections à large échelle.

Deuxièmement, la généralisabilité de ces approches neuronales peut être rapidement impactée dans ce contexte limité en données. En effet, il est courant de rencontrer deux défis majeurs lors de l'implémentation de ces approches : un changement de distribution, aussi appelé domaine, ou un surapprentissage. Les changements de domaine surviennent dès lors que la distribution relative aux données n'est plus la même entre l'entraînement de l'algorithme et son application pratique. Typiquement, un algorithme entraîné sur des images radar en milieu urbain ne sera pas généralisable directement à des images radar en milieu montagneux - notamment à cause des différences de géométrie, de couverture du sol et de distorsions géométriques. La problématique du surapprentissage survient lorsque la configuration d'un entraînement est inadaptée aux

données, au modèle et à la tâche à effectuer. Ceci a pour effet de spécialiser le modèle neuronal uniquement sur les données d'entraînement si bien que toute nouvelle donnée sera traitée de façon aléatoire par le réseau de neurones. De fait, la problématique de généralisabilité est cruciale en apprentissage profond, et plus encore dans un contexte de données limitées et variables comme ce peut être le cas de la détection d'instabilités gravitaires.

Enfin, le milieu des géosciences étant constamment à la recherche de la compréhension des phénomènes géophysiques, il est nécessaire que tout résultat obtenu soit étayé par des preuves pour pouvoir être validé. L'une des critiques majeures faite à l'encontre des approches d'apprentissage automatique est leur manque d'explicabilité et de compréhension. Ceci leur vaut très souvent le qualificatif de "boîte noire" indiquant qu'il suffit de leur fournir une donnée d'entrée pour obtenir "de façon magique" un résultat en sortie. Ainsi donc, mettre en place des approches neuronales pour la détection d'instabilités gravitaires nécessite à la fois une prise de décision fiable mais aussi des explications rendant ces décisions crédibles et compréhensibles.

Problématiques

Ce rappel contextuel relatif aux trois thématiques de recherche composant cette thèse a permis de mettre en lumière les grandes problématiques liées à chacune. Ainsi donc, avant même de chercher à détecter des instabilités gravitaires, une question relative aux données se pose :

Comment utiliser et interpréter des images SAR contenant des distorsions géométriques pour remplacer les images optiques couvertes de nuages ?

Afin de répondre à cette problématique, le premier axe de recherche étudié porte sur la traduction de modalités. L'objectif d'une telle traduction et de transformer une image SAR contenant des distorsions géométriques en une image optique interprétable et sans artefacts. Ainsi donc, un nouveau réseau de neurones, appelé SARDINet (pour *SAR Distorted Image translator Network* en anglais), est proposé dans ces travaux pour assurer cette traduction SAR-optique. Plusieurs contributions supplémentaires ont également été mises en place pour améliorer la pertinence des motifs optiques reconstruits ainsi que leur contraste et leur niveau de détail.

Mais même si les résultats de traduction peuvent *paraître* correctes quantitativement et qualitativement, la question de leur exploitabilité pour mener à bien la détection d'instabilités gravitaires reste entière. L'objectif de cette traduction étant de substituer les images optiques inutilisables pour cause de couverture nuageuse lors du déclenchement soudain de glissements de terrain, il est donc nécessaire qu'elle soit aussi adaptée à la reconstruction des motifs spécifiques à ces phénomènes. En effet, les auteurs de [Wang et al., 2022a] ont démontré que la complexité de la couverture du sol influençait significativement la qualité de la traduction SAR-optique. De plus, la représentativité des motifs propres aux instabilités gravitaires est très faible, voire inexistante, dans les jeux de données standards utilisés pour la traduction de modalité (*e.g.* BigEarthNet-MM [Sumbul et al., 2021], SEN12 [Schmitt et al., 2018], SpaceNet6 [Shermeyer et al., 2020]). Ainsi donc, l'utilisation de la traduction de modalité SAR-optique pour la détection d'instabilités gravitaires pose une autre question :

Comment guider la traduction de modalité pour reconstruire des images optiques contenant de véritables motifs de glissements exploitables par des modèles de détection de changements ?

Cette problématique est d'autant plus complexe que chaque tâche visée peut avoir ses propres exigences quant aux caractéristiques présentes sur l'image. Ainsi donc, après avoir testé des approches de détection de changements spécialisées dans les glissements de terrain et entraînées sur des images traduites, il apparaît que les motifs reconstruits ne sont pas suffisamment caractéristiques de la tâche finale. Deux stratégies de guidage sont alors proposées. La première intitulée TCL (pour *Translation-Classification Loss* en anglais) permet de guider la reconstruction des caractéristiques optiques à l'aide d'une classification post-traduction. Cette classification est ensuite remplacée par une détection de changements pour forcer le réseau à reconstruire des motifs utiles à cette tâche, et en particulier pour la détection de glissements de terrain à l'aide de la TCDL (acronyme anglais pour *Translation-Change Detection Loss*).

Enfin, la thèse visant la détection d'instabilités à la fois soudaines et lentes, des travaux ont été menés pour adapter les algorithmes d'apprentissage automatique au traitement d'interférogrammes. Mais si la traduction SAR-optique permettait de ramener de l'interprétabilité dans les décisions prises par les réseaux de neurones, l'exploitation directe d'interférogrammes est beaucoup plus limitée en terme de compréhensibilité. De plus, les interférogrammes peuvent être sujets à différentes perturbations causant la génération de motifs contrefactuels à cause des zones de distorsion et des phénomènes atmosphériques et topographiques. Ainsi donc leur exploitation fait émerger une nouvelle problématique :

Comment identifier et justifier de la présence d'instabilités gravitaires à mouvement lent sur des interférogrammes ?

Afin de mieux cerner ces phénomènes sur de telles données, une marche à suivre a été établie en collaboration avec des géomorphologues pour assurer des identifications fiables. Cette marche à suivre a été appliquée sur plusieurs interférogrammes dans les Alpes françaises afin de créer un nouveau jeu de données, ISSLIDE (pour *Interferogram dataset for Slow SLIding area DEtection* en anglais), adapté aux approches d'apprentissage profond et annoté manuellement pour assurer la fiabilité des détections. Suite à la création et l'exploitation de ce jeu de données par des modèles de segmentation standards, une dernière contribution est proposée afin de pouvoir justifier de façon efficace les décisions prises par un réseau de neurones. De cette façon, une nouvelle stratégie, baptisée ECSPLAIN (pour *Explainability-Constrained claSsifier for Pairing the detection and Localization of slow moving Areas on INterferograms* en anglais), intègre des méthodes d'explicabilité durant l'entraînement d'un classificateur. Ainsi, ce dernier a non seulement pour objectif de détecter la présence ou l'absence d'instabilité sur une imagette d'interférogramme, mais aussi de justifier cette décision en générant une carte d'activation (appelé CAM pour *Class Activation Maps* en anglais) correspondant à la segmentation de l'instabilité en question.

Ainsi donc, du fait des trois grandes problématiques adressées dans cette thèse, la suite de ce document se structure en trois sections successives décrivant les contributions en traduction de modalité SAR-optique, celles visant à rendre les traductions utilisables dans le cadre de la détection de glissements de terrain soudains et enfin celles visant à exploiter les interférogrammes pour la détection d'instabilités gravitaires lentes. La toute dernière section détaille les conclusions et perspectives issues de ces travaux de thèse.

Traduction de Modalité SAR-optique

La vaste majorité des modèles d'apprentissage profond ayant pour objectif la transformation d'une image en une autre est basée sur des réseaux adversaires génératifs (GAN en anglais) [Goodfellow et al., 2016]. Les approches de traduction de modalité ne font pas exception et sont généralement modélisées selon l'architecture Pix2pix [Isola et al., 2017] ou CycleGAN [Zhu et al., 2017] en fonction du niveau de supervision requis. Ces approches ont rapidement été étendues à la télédétection et à la traduction SAR-optique comme en témoignent les nombreux travaux s'y référençant [Lee and Park, 2023, Yang et al., 2022b, Zhang et al., 2020b, Fu et al., 2021, Luppino et al., 2021].

Mais deux caractéristiques propres aux GANs sont tout de même à mentionner et tempèrent leur utilisation. Premièrement, de part la compétition entre le générateur et le discriminateur, l'hyperparamétrisation de l'entraînement est extrêmement sensible et les résultats peuvent être très instables d'un entraînement à l'autre. Deuxièmement, les approches génératives ont pour objectif premier de construire des images *ressemblant* à des images réelles. Ceci signifie donc que les motifs reconstruits sont avant tout générés dans le but d'être crédibles. Ainsi le réseau générateur est fortement susceptible de créer des motifs propres à des images optiques mais sans avoir la moindre preuve qu'ils correspondent réellement à ce qui est visible dans une image SAR. En particulier, dans le cadre d'une zone de distorsion géométrique, peu sinon aucune information n'est disponible. Un réseau générateur peu alors avoir tendance à halluciner des motifs, ce qui pourraient mener à des interprétations optiques non conformes à la réalité de terrain. Ceci est vérifié quantitativement dans [Liu et al., 2021a] où un entraînement adversaire réduisait significativement les métriques révélatrices de la pertinence des motifs reconstruits.

Ceci a donc poussé plusieurs auteurs à contraindre ou à s'affranchir de l'entraînement adversaire pour éviter ces problèmes et à trouver des alternatives plus fiables. Ainsi, des approches basées sur la décomposition en ondelettes [Wei et al., 2023a, Li et al., 2022b], des *transformers* [Wang et al., 2022d] ou des architectures plus spécifiques [Luppino et al., 2024, Xia et al., 2024] ont été mises en place et ont démontré des résultats compétitifs avec les approches génératives. Les travaux menés dans cette thèse s'intègrent également dans cette dynamique. Pour assurer des traductions avant tout pertinentes, le choix a été fait de développer de nouvelles approches sans considérations adversaires. Un tel apprentissage n'est considéré qu'en dernier recours pour des fins de visualisation.

Traduire malgré les distorsions

La première contribution majeure de cette thèse réside dans l'implémentation d'un réseau de neurones traducteur d'images SAR souffrant de distorsions géométriques nommé SARDINet de part son acronyme anglais (voir Section 4.3) et proposé dans [Bralet et al., 2022]. Pour ce faire, l'architecture encodeur-décodeur disponible en Figure 4.2 se divise en trois étapes différentes :

- Une étape de conditionnement composée de simples convolutions 2D permettant de filtrer le bruit de chatoiement et d'extraire les premières caractéristiques évidentes de l'image SAR d'entrée.
- Un encodeur multi-échelle composé de trois branches indépendantes et dédiées à une résolution spécifique. En effet, le traitement appliqué par chaque branche est plus ou moins poussé en fonction de la résolution des caractéristiques courantes. Ainsi, la première branche permet une extraction globale quand la seconde et troisième branche permettent une extraction spécifique à basse résolution et haute résolution. De plus, chaque branche est composée de blocs de convolutions séparables [Chollet, 2017]. Ces blocs sont constitués de deux convolutions successives : une convolution spatiale et une convolution inter-canal. La première consiste à appliquer à chaque canal d'entrée un noyau de convolution spécifique et à concaténer les cartes d'activation résultantes. La seconde consiste en une combinaison linéaire des cartes d'activation précédentes à l'aide d'une convolution 1D appliquée sur la dimension des canaux.
- Un décodeur contenant plusieurs branches pour la reconstruction de chaque canal de sortie indépendemment les uns des autres. Ce décodeur rend l'architecture particulièrement flexible quant au nombre de canaux de sortie à reconstruire.

L'entraînement de SARDINet, Pix2pix [Isola et al., 2017] et CycleGAN [Zhu et al., 2017] est effectué sur un sous-ensemble du jeu de données[5] SpaceNet 6 [Shermeyer et al., 2020]. Ce sous-ensemble de couples SAR-optique acquis à Rotterdam contient des images avec de fortes distorsions géométriques en milieu urbain (voir Figure 2.10). Qu'elles soient quantitatives ou qualitatives, les performances de SARDINet sur ce jeu de données sont démontrées comme plus pertinentes du point de vue structurel et sémantique. En effet, les reconstructions de Pix2pix et CycleGAN sont certes plus contrastées mais contiennent de nombreux artefacts qui ne permettent pas une utilisation fiable *a posteriori* de ces traductions.

Pertinence des traductions

Deux études complémentaires ont été menées afin d'améliorer les performances de SARDINet en milieu rural et montagneux.

La première étude détaillée dans [Bralet et al., 2024c] consiste à fournir le Modèle Numérique de Terrain (DEM en anglais) en entrée de SARDINet. En effet, le DEM permettant de recenser l'altitude de chaque pixel considéré, il est intimement lié aux pentes qui causent des distorsions géométriques sur les images SAR. L'étude visait à trouver la meilleure stratégie de fusion de ce DEM et des images SAR pour tirer le meilleur parti des deux données d'entrée. Ainsi, trois stratégies de fusion ont été implémentées au niveau de l'étape de conditionnement de SARDINet (laissant l'encodeur multi-échelle et le décodeur inchangés) comme détaillé en Figure 4.4.

[5] https://ieee-dataport.org/documents/can-artificial-intelligence-untangle-distorted-and-compressed-geometries-associated-sar, dernière visite le 28 juin 2024.

- La fusion précoce (SARDINet$_E$) consiste à concaténer le DEM et l'image SAR dual-polarisation pour former une donnée à trois canaux d'entrée. Le réseau reste donc inchangé, excepté la première couche qu'il a fallut adapter à cette nouvelle dimension.

- La fusion tardive (SARDINet$_L$) consiste à traiter chaque image dans une branche indépendante puis concaténer les caractéristiques extraites.

- La fusion progressive (SARDINet$_P$) reprend la même idée que la fusion précoce mais remplace les couches de convolution 2D de l'étape de conditionnement par des blocs de convolution séparable. De cette façon, le réseau peut extraire successivement des caractéristiques spatiales séparément sur chaque canal puis les combiner. Cette approche permet aussi de réduire significativement le nombre de paramètres dans le réseau de neurones.

Ces trois différents réseaux sont comparés à SARDINet entraîné sans le DEM et aux deux réseaux adversaires Pix2pix [Isola et al., 2017] et CycleGAN [Zhu et al., 2017]. Puisque cette fusion a un intérêt essentiellement dans des zones de distorsions géométriques, un nouveau jeu de données multimodal en milieu montagneux est également proposé pour effectuer les tests et est mis à disposition sur IEEE Dataport[6]. Les résultats de chaque traducteur appliqué sur ce nouveau jeu de données a permis de statuer quant à la pertinence du DEM lors des traductions de modalité en milieu montagneux puisqu'il permet de limiter des confusions, notamment entre les zones d'ombre radar et celles relatives aux plans d'eau. De plus, les résultats renforcent les conclusions précédemment obtenues avec SARDINet en démontrant que cette architecture permet d'exploiter intelligemment l'image SAR et le DEM contrairement aux approches adversaires à la fois instables durant l'entraînement et non fiables dans leurs reconstructions à cause des artefacts hallucinés.

La seconde étude proposée et détaillée dans [Bralet et al., 2023] vise à comparer différentes stratégies de décodage utilisées pour sur-échantillonner les caractéristiques latentes de SARDINet. Trois stratégies illustrées en Figure 4.7 sont implémentées dans ces travaux :

- L'interpolation (SARDINet$_{up}$) consiste à utiliser une méthode d'interpolation classique (plus proche voisin, bilinéaire, bicubique, etc.) pour effectuer la mise à l'échelle.

- La convolution transposée (SARDINet$_{tr}$) consiste à disperser les pixels basse résolution sur une grille haute résolution et à combler les interstices en appliquant une convolution sur la grille haute résolution. De cette façon, le réseau est libre de déterminer sa propre interpolation.

- La convolution sub-pixellique (SARDINet$^{pix}_{pix}$) proposée dans [Shi et al., 2016] consiste à déterminer cette interpolation avant la mise à l'échelle. Ainsi, plusieurs convolutions sont appliquées aux caractéristiques basse résolution si bien qu'une simple réorganisation des activations permet de générer une image de plus haute résolution.

Les couches de convolution sub-pixellique étant initialement proposées à des fins de super-résolution, deux stratégies supplémentaires sont considérées : SARDINet$^{pix}_{tr}$ et SARDINet$^{pix}_{up}$. Chacune d'elle n'utilise une convolution sub-pixellique que dans la toute dernière couche de sur-échantillonnage alors que les couches précédentes sont assurées respectivement par des convolutions transposées ou des interpolations.

L'application de ces cinq stratégies à un grand jeu de données standard de la littérature, BigEarthNet-MM [Sumbul et al., 2021], a permis de conclure que l'utilisation des convolutions sub-pixelliques avaient un réel intérêt lorsqu'elles étaient appliquées en toute dernière couche et que les interpolations renforçaient le flou présent dans l'image reconstruite. Ainsi, d'après ces travaux, la meilleure stratégie à considérer est SARDINet$^{pix}_{tr}$, composée de convolutions transposées dans les couches les plus basses et d'une convolution sub-pixellique en toute dernière couche.

[6] https://ieee-dataport.org/documents/sar-dem-optical-mountainous-dataset-distortion-management, dernière visite le 28 juin 2024

Contraste des traductions

Malgré que les deux études précédentes ont démontré de réelles améliorations en terme de perspicacité et pertinence des traductions, celles-ci manquent toujours d'un contraste certain. De fait, quatre études ont été menées afin de renforcer le contraste des traductions. Il est important de mentionner ici que le renforcement du contraste ne doit en aucun cas impacter la pertinence des motifs reconstruits, auquel cas la méthode serait considérée comme utile "visuellement" mais moins exploitable *a posteriori*.

La première modification proposée dans [Bralet et al., 2024b] est architecturale. En effet, SARDINet a été modifié en SARDINetV2 visible en Figure 4.9 en ajoutant des connections entre l'encodeur et le décodeur. De cette façon, les détails structurels extraits des images SAR peuvent être transférés bien plus facilement au décodeur ce qui permet de rehausser le contraste des traductions. D'autres modifications ont également été apportées telles qu'un rééquilibrage des poids du réseau, un traitement avancé au niveau de l'espace latent et une fusion apprise des caractéristiques extraites dans l'encodeur multi-échelle. Des expériences menées sur SpaceNet 6 et BigEarthNet-MM ont permis de conclure que SARDINetV2 améliore significativement les performances initiales de SARDINet. Non seulement le contraste des images a été largement rehaussé, mais la pertinence des motifs reconstruits a également été améliorée pour une meilleure interprétation des zones d'étude. De plus, les performances en milieu rural sur des images de moyenne résolution de BigEarthNet-MM (Sentinel-1 et Sentinel-2 à 10m) ont permis de prouver que SARDINetV2 compensait les lacunes de SARDINet sur ces images plus difficilement interprétables que les images urbaines aéroportées de SpaceNet 6.

Une seconde expérience menée dans [Bralet et al., 2022] consiste en un entraînement adversaire de SARDINet à l'aide d'un discriminateur très simple pour limiter le risque d'hallucination du réseau. Néanmoins, cet entraînement adversaire conduit sur SpaceNet 6 n'a pas donné de résultats concluants. En effet, les résultats étaient à la fois moins bons quantitativement parlant mais également beaucoup plus perturbés par les distorsions géométriques et les objets fortement réflecteurs d'un point de vue qualitatif. Ainsi donc, même si le contraste a effectivement été rehaussé, la perte en pertinence des motifs reconstruits ne permet pas de considérer un entraînement adversaire de SARDINet comme fiable pour la traduction SAR-optique.

Une troisième possibilité considérée dans [Bralet et al., 2024c] consiste à appliquer des méthodes de rehaussement de contraste classiques de traitements d'image aux traductions. Ainsi, un filtrage de Laplace et un filtrage de Wiener ont été appliqués aux images traduites afin d'observer le gain en contraste et l'évolution des métriques liés à ces traitements. D'après les performances observées sur BigEarthNet-MM, chacun des filtres apporte un réel gain sur le contraste final de l'image et ce avec un impact modéré sinon inexistant sur la pertinence des motifs reconstruits. Certes ces reconstructions ne permettent pas d'égaler les performances de SARDINetV2 mais elles montrent que le flou peut être facilement compensé *a posteriori* avec des algorithmes standards.

Enfin, la dernière proposition consiste à combiner les trois expériences précédentes en une stratégie de Traduction-Défloutage (Tra-De pour *Translation-Deblurring* en anglais). Celle-ci vise à appliquer SARDINetV2 pour effectuer la traduction de modalité SAR-optique initiale puis à entraîner un réseau adversaire de la littérature, Pix2pix, à déflouter l'image *a posteriori*. Les premiers résultats appliqués sur BigEarthNet-MM ont permis de démontrer que malgré une légère perte en pertinence des motifs, le contraste est bien meilleur et l'image est d'autant plus compréhensible. Cette première expérience est particulièrement encourageante puisqu'en comparaison avec les approches adversaires de la littérature, elle permet tout de même de conserver les informations caractéristiques issues de la traduction de modalité.

Détection de Glissements de Terrain

Suite aux performances en traduction de modalité observées avec l'implémentation et les améliorations de SARDINet, l'objectif est désormais de vérifier l'applicabilité de ces traductions dans un contexte de détection de changements spécialisé dans les glissements de terrain soudains.

La majorité des approches de détection de changements se base sur un réseau de type Siamois [Bromley et al., 1993] visible en Figure 3.11. Cette architecture a pour avantage d'utiliser un seul encodeur pour traiter les deux images pre- et post-changement. Ainsi, une simple comparaison des caractéristiques extraites permet

de trouver les changements majeurs survenus entre les deux images. Cette simple comparaison explique son utilisation dans de nombreux domaines, y compris en télédétection [Ebel et al., 2021, Amankwah et al., 2022].

Si les détections de changements monomodales ont été largement implémentées pour une meilleure compréhension de l'utilisation des sols [Fang et al., 2023, Adriano et al., 2020, Cambrin et al., 2023], les approches multimodales sont plus rares. En particulier, seules quelques approches proposent de coupler la traduction de modalité avec la détection de changements [Luppino et al., 2021, Li et al., 2021b, Luppino et al., 2024, Zhao et al., 2023]. Au sein de cette thèse, nous proposons d'enrichir ces recherches en utilisant des réseaux de détection standards, entre autre un réseau Siamois, appliqués sur les traductions. Afin de renforcer la compréhension du sol par le réseau de traduction, deux nouvelles stratégies sont proposées combinant la tâche de traduction de modalité avec une tâche auxiliaire appliquée directement aux images traduites durant l'entraînement du traducteur.

Détection sur les traductions

Une toute première étude implémente trois réseaux légers de détection de changements afin d'étudier leur efficacité dans trois configurations différentes : une détection monomodale optique, une détection monomodale SAR et une détection multimodale optique-traduction. L'objectif de l'étude courante est évidemment d'obtenir une détection multimodale optique-traduction plus performante que la monomodale SAR permettant de prendre de premières mesures d'urgence dans l'attente d'une image optique post-évènement exploitable.

Ainsi donc trois réseaux de type U-Net [Ronneberger et al., 2015] avec des blocs résiduels [He et al., 2016a] ont pu être implémentés comme illustré en Figure 5.3 :

- CDNet$_E$ où les images pre- et post-évènement sont concaténées en entrée du réseau,
- CDNet$_L$ où chaque image est traitée par un encodeur indépendant,
- CDNet$_S$ où chaque image est traitée indépendamment par le même encodeur comme pour un réseau Siamois.

Malgré la disponibilité du jeu de données Landslide4Sense [Ghorbanzadeh et al., 2022a], ce dernier ne permettait ni une approche bitemporelle, ni une approche multimodale. De fait, un nouveau jeu de données a été créé afin de palier cette problématique. Le jeu de données rassemble des paires optiques et des paires SAR d'images. Chaque paire est composée d'une image avant et après le séisme de magnitude 7.2 qui a touché Haïti le 14 août 2021. Ce jeu de données est mis à disposition sur IEEE Dataport[7] pour permettre de plus amples études sur des glissements de terrain dans des conditions météorologiques, de végétation, de résolution et de distorsions géométriques difficiles.

Suite à un raffinage de SARDINetV2 sur ce nouveau jeu de données et à l'application des trois CDNets dans les trois configurations, une conclusion évidente est apparue : les traductions n'ont pas permis d'améliorer les performances monomodales SAR. En effet, chaque CDNet souffre d'une sur-segmentation des glissements de terrain ne permettant pas une détection fiable et efficace des phénomènes. Plusieurs possibilités peuvent expliquer une telle contre performance : l'architecture des CDNet, la complexité du jeu de données ou encore le manque de pertinence des traductions par rapport à la tâche d'intérêt.

Mieux comprendre le terrain...

Suite à cette étude et aux conclusions extraites de [Wang et al., 2022a] expliquant que la couverture du sol pouvait influencer la qualité et la compréhensibilité des traductions, les travaux de thèse se sont concentrés sur l'amélioration des performances de traduction de modalité.

Ainsi, pour mieux intégrer la compréhension de la couverture du sol dans le processus de traduction de modalité, une nouvelle fonction de coût est proposée dans [Bralet et al., 2024b] combinant une tâche

[7] https://ieee-dataport.org/documents/multimodal-remote-sensing-dataset-landslide-change-detection-haiti, dernière visite le 16 juillet 2024.

de traduction et une tâche de classification (nommée TCL en anglais). Cette fonction de coût se base sur l'utilisation d'un classificateur ResNet50 [He et al., 2016a] indépendant et pré-entraîné sur des images optiques réelles pour évaluer la qualité des traductions reconstruites par SARDINetV2. Ce classificateur vise à prédire les différentes classes composant la couverture du sol sur chaque image (village, forêt, fleuve, champs, etc.). L'exploitation du classificateur se fait en gelant l'ensemble de ses poids durant l'entraînement du traducteur. Ainsi, quatre termes assurent la composition de la TCL :

- L'erreur pixellique absolue (*Mean Absolute Error* ou MAE en anglais) entre la vérité optique et l'image traduite.
- L'erreur en terme de similarité structurelle (*Structural Similarity Index Measure* ou SSIM en anglais) entre la vérité optique et l'image traduite.
- L'erreur absolue calculée entre les caractéristiques extraites par les couches cachées du classificateur à partir de la vérité optique et de l'image traduite.
- L'erreur d'entropie croisée calculée entre les prédictions du classificateur appliqué à l'image traduite et les véritables étiquettes associées à chaque image.

Cette TCL a été utilisée pour l'entraînement de 6 traducteurs non-adversaires : Pix2pix, CycleGAN, GCGAN [Fu et al., 2019], SRM [Zhang et al., 2022], SARDINet et SARDINetV2. En effet, chaque approche adversaire a vu son discriminateur désactivé durant l'entraînement pour éviter la génération d'artefacts qui nuiraient à l'interprétabilité de la TCL. Suite aux expériences conduites sur BigEarthNet-MM et comparées aux entraînements de ces mêmes réseaux avec leur fonction de coût "traditionnelle", un constat sans ambiguïté a pu être établi : la TCL a un réel impact sur le contraste, la pertinence et la classificabilité des images traduites. Une étude d'ablation a également été menée afin de mieux comprendre l'impact de chaque terme sur les reconstructions. Ainsi donc, si la MAE est prise comme fonction de coût standard, l'utilisation de la SSIM permet une nette amélioration du contraste des images quand la comparaison des caractéristiques extraites par le classificateur permet d'obtenir des détails plus intéressants encore et de combler les lacunes et confusions de la SSIM seule. Le dernier terme d'entropie croisée a essentiellement un impact global sur les couleurs et la luminosité d'une classe par rapport à une autre. Un point particulièrement intéressant de cette étude est l'applicabilité de la TCL même dans le cas où les étiquettes de couverture du sol ne seraient pas disponibles. En effet, les résultats ont démontré que les trois premiers termes de la TCL permettaient d'obtenir les traductions les plus pertinentes et contrastées malgré une légère perte de classificabilité. Enfin, une étude complémentaire a permis de voir que l'utilisation de la TCL lors d'un entraînement adversaire permettait de régulariser et stabiliser les performances tout en limitant, sans les exclure, la génération d'artefacts dans les images reconstruites.

... Pour mieux détecter les glissements soudains

Suite aux performances particulièrement encourageantes de la TCL, nous proposons ici de l'étendre à une autre tâche : la détection de changements. Ainsi donc, la TCL est modifiée en une fonction de coût couplant la tâche de traduction de modalité et de détection de changements induits par les glissements de terrain (TCDL en anglais). Le principe et les termes mentionnés précédemment restent les mêmes tandis que le classificateur est remplacé par le réseau Siamois CDNet_S entraîné sur le jeu de données d'Haïti détaillé plus haut. Même si CDNet_S ne montre pas les meilleures performances de détection de changements, il permet de faciliter grandement l'implémentation et l'apprentissage de SARDINetV2 pour deux raisons. Premièrement, les images sont traitées indépendamment l'une de l'autre ce qui permet de mieux identifier les caractéristiques extraites propres à chaque image contrairement à CDNet_E où les images passent ensemble dans l'encodeur. Deuxièmement, de part le caractère partagé de l'encodeur, les deux images pre- et post-évènements sont utilisées pour entraîner les mêmes poids ce qui permet de renforcer l'extraction de caractéristiques propres aux images optiques contrairement à CDNet_L.

Au delà des considérations architecturales, les termes de la fonction de coût de MAE et de SSIM restent inchangés. Concernant le terme de comparaison de caractéristiques, seules les caractéristiques de l'encodeur

sont comparées dans ces travaux, en effet, le décodeur traitant la fusion des deux images d'entrée, il ne permet pas d'évaluer uniquement la qualité d'une image en particulier. Concernant le terme d'entropie croisée, celui ci n'est plus appliqué sur un vecteur assimilable à la détection de plusieurs classes de couverture du sol mais directement sur la carte de détection de changements.

Ainsi donc, en partant du modèle SARDINetV2 pré-entraîné avec la TCL sur BigEarthNet-MM comme présenté plus haut, et en le raffinant avec la TCDL sur le jeu de données d'Haïti, de nouvelles traductions ont pu être générées montrant un meilleur contraste et un renforcement des zones affectées par les glissements de terrain. Malgré ces améliorations significatives comparées aux images traduites initiales, l'application des différents CDNet n'a pas permis d'atteindre des performances de détection de changements meilleures que celles de la configuration monomodale SAR. En effet, le phénomène de sur-segmentation est toujours relativement présent sur les toutes premières expériences menées. Une nette amélioration a tout de même pu être notée au sein des résultats de détection comparés aux résultats initiaux avec un bien meilleur équilibre entre la précision et le rappel laissant présager que des expériences plus approfondies sur le sujet pourraient à terme être fructueuses.

Détection d'Instabilités à Mouvement Lent

Si la détection du déclenchement des instabilités gravitaires permet de réagir au plus vite lors d'un tel évènement et secourir un maximum de personnes, la détection des instabilités susceptibles de se déclencher est un sujet tout autre qui permet d'anticiper les dégâts. Mais pour pouvoir prédire de tels déclenchements, encore faut-il connaître l'emplacement de ces instabilités gravitaires. C'est dans ce contexte que se placent les dernières contributions proposées dans cette thèse. En effet, l'objectif ici est de faire une cartographie automatique des zones en mouvement pour que de futurs algorithmes puissent ensuite étudier l'évolution de ces zones de mouvement et prédire les conditions pour leur déclenchement.

Cependant, les images satellitaires d'une résolution de 10 mètres telles que Sentinel-2 ne permettent pas de détecter des mouvements de quelques centimètres par an. Ainsi donc, la voie privilégiée pour une telle étude reste l'utilisation d'interférogrammes à partir d'images SAR. La volcanologie notamment est particulièrement avancée dans l'utilisation d'interférogrammes pour détecter les éruptions volcaniques [Anantrasirichai et al., 2019, Biggs et al., 2021, Bountos et al., 2022b]. Ces interférogrammes ont également été exploités pour une cartographie des zones de mouvement gravitaire [Bodin, 2020, Bertone et al., 2022, Rouyet et al., 2021], mais très peu d'études ont été menées pour combiner à la fois les approches d'apprentissage automatique et la détection d'instabilités gravitaires sur des interférogrammes.

Un tel constat est également dû au manque de compréhensibilité des algorithmes d'apprentissage automatique. En effet, additionné à l'interprétation difficile des interférogrammes et à la variabilité des phénomènes géophysiques étudiés, un manque de confiance dans les détections peut aisément apparaître. C'est pour cette raison que les quelques approches implémentant des algorithmes automatiques dans de telles conditions proposent d'expliquer leurs détections à l'aide de méthodes standards d'explicabilité et en particulier avec des cartes d'activation (CAM) [Brengman and Barnhart, 2021, Zhao et al., 2022b, Chen et al., 2024].

Une carte d'activation est une approche d'explicabilité permettant de générer une carte de chaleur indiquant les zones qui ont le plus contribué à la décision finale prise par le réseau neuronal. Tout un état de l'art existe sur les cartes d'activation, qu'elles soient basées sur un calcul de gradient [Selvaraju et al., 2017, Jiang et al., 2021, Rebuffi et al., 2020], sur des masques [Wang et al., 2020a, Naidu et al., 2020, Desai and Ramaswamy, 2020], sur plusieurs réseaux parallèles [Zhang et al., 2018b, Chen et al., 2022b] ou sur des méthodes plus sophistiquées encore [Wang and Bunyak, 2023, Muhammad and Yeasin, 2020]. Si ces CAMs sont couramment utilisées pour servir de pseudo-labels lors d'entraînements faiblement supervisés [Miao et al., 2023, Liu and Zhang, 2023], rien ne garanti de leur fiabilité.

Elles ont en effet démontré des résultats particulièrement variables en fonction des couches neuronales expliquées [Rebuffi et al., 2020] voire même de la tâche implémentée. C'est ainsi que l'implémentation de méthodes contraignant l'explicabilité des réseaux ont pu émerger [Ross et al., 2017, Wargnier-Dauchelle et al., 2023, Boyd et al., 2023]. Une telle contrainte force le réseau à prendre des décisions basées sur des preuves concrètes et des critères nécessaires à la fiabilité des résultats.

Les travaux finaux menés dans cette thèse visent donc à exploiter ces méthodes d'explicabilité contrainte afin de mieux pouvoir interpréter et comprendre les décisions prises par les réseaux de neurones quant à la détection de zones en mouvement sur des interférogrammes.

Identifier les instabilités

Malgré tout, avant de pouvoir détecter de façon fiable des instabilités gravitaires à l'aide d'approches automatiques, encore faut-il savoir les identifier visuellement sur des interférogrammes.

Cette problématique a conduit à une collaboration avec des géomorphologues des laboratoires EDYTEM et ISTerre afin d'établir une routine d'identification composée de plusieurs critères :

- Identifier des franges ou des bulles de différence de phase telles que sur la Figure 6.2,
- S'assurer que la détection n'a pas lieu dans une zone de distorsion géométrique (ininterprétable),
- S'assurer que la cohérence de l'interférogramme est suffisamment forte pour que la différence de phase soit fiable,
- S'assurer de la répétitivité du mouvement sur différents interférogrammes,
- Optionnellement, vérifier que le mouvement est visible sur des interférogrammes avec différentes bases de temps (*e.g.* 6, 12 et 18 jours),
- Rechercher des indices géomorphologiques sur des images optiques.

La prise en compte de l'ensemble de ces critères a permis la création d'un nouveau jeu de données adapté aux approches d'apprentissage automatique, ISSLIDE en anglais. Ce jeu de données, contenant les annotations manuelles de 200 instabilités gravitaires identifiées sur 54 interférogrammes, est décrit en détails dans [Bralet et al., 2024d]. Il est également disponible sur IEEE Dataport[8] sous deux formats : un jeu de données brut et un jeu de données prêt à utilisation. Le jeu de données brut contient l'ensemble des interférogrammes calculés à 6, 12 et 18 jours ainsi qu'un fichier d'annotation spécifique à chaque interférogramme. Le jeu de données prêt à l'utilisation contient des imagettes d'interférogramme sur les zones identifiées comme "en mouvement" pour pouvoir être directement utilisé par des approches d'apprentissage automatique.

Afin de pouvoir vérifier l'exploitabilité de ce nouveau jeu de données, plusieurs réseaux de neurones de segmentation standards ont été entraînés tels que FCN [Long et al., 2015], DeepLabV3 [Chen et al., 2017] et des dérivés de U-Net [Ronneberger et al., 2015] avec des blocs résiduels [He et al., 2016a] ou des convolutions séparables [Chollet, 2017]. Les résultats obtenus par chaque réseau permettent de conclure quant à la difficulté de la tâche mais également à sa faisabilité. En effet, les performances obtenues sont certes relativement satisfaisantes mais laissent tout de même une marge d'amélioration pour cette tâche de segmentation. De plus, les poids des réseaux de neurones entraînés sur le jeu de données PASCAL VOC [Everingham et al., 2010] ne sont pas transférables aux données d'interférogrammes comme le montre la baisse des performances observée suite au raffinage des réseaux pré-entraînés. Ainsi, ces résultats permettent de démontrer que la segmentation d'instabilités gravitaires sur des interférogrammes reste un sujet ouvert à de plus amples recherches au vue des performances des approches standards de la littérature.

Détecter et expliquer

La contribution finale proposée dans cette thèse a pour objectif de répondre aux exigences de compréhensibilité des géosciences, de légèreté et d'efficacité des approches neuronales, et de généralisabilité des images satellitaires. En effet, cette contribution finale détaillée dans [Bralet et al., 2024a] consiste en la mise en place d'un classificateur dont l'explicabilité est contrainte afin de coupler la détection et la localisation des zones

[8] https://ieee-dataport.org/documents/isslide-insar-dataset-slow-sliding-area-detection-machine-learning, dernière visite le 24 juillet 2024.

en mouvement sur des interférogrammes (ECSPLAIN en anglais). Cette approche vise tout d'abord à assurer l'explicabilité du réseau de neurones, à entraîner ce classificateur à effectuer deux tâches distinctes à partir d'une architecture unique, et à être généralisable à différentes variations d'interférogrammes et de mouvements gravitaires.

Pour ce faire, un classificateur ResNet50 est premièrement implémenté et modélisé afin de répondre à une tâche unique : détecter la présence d'un mouvement de terrain sur une imagette d'interférogramme. Suite à chaque prédiction, la stratégie ECSPLAIN consiste à calculer une CAM afin de justifier d'une telle décision. La particularité de la stratégie ECSPLAIN est que cette CAM n'est pas générée **suite** à l'entraînement mais bien **durant** l'entraînement du classificateur comme décrit dans la Figure 6.8. De cette façon, en utilisant une fonction de coût basée sur la justesse des prédictions ainsi que sur la qualité de l'explication par rapport à la véritable annotation, le réseau classificateur est entraîné à prendre des décisions en identifiant des caractéristiques propres à la localisation du mouvement effectif.

Cette stratégie est particulièrement utile au temps d'inférence puisque le calcul de la carte de localisation peut être restreint aux imagettes où un mouvement a été détecté. De la même façon, l'utilisateur final peut toujours demander à calculer la carte de localisation pour afficher l'emplacement le plus probable d'un mouvement, et ce même dans le cas où un mouvement aurait été manqué durant la phase de détection du réseau. Une telle stratégie permet donc à l'utilisateur de garder la main sur les décisions de l'algorithme, de le contester voire même de le corriger si le besoin s'en fait ressentir.

Pour s'assurer de la qualité des décisions et des cartes de localisation calculées par le classificateur entraîné avec la stratégie ECSPLAIN, plusieurs expériences ont été menées sur le jeu de données prêt à utilisation d'ISSLIDE :

- Plusieurs CAMs ont été implémentées : Grad-CAM [Selvaraju et al., 2017], Layer-CAM [Jiang et al., 2021], NormGrad [Rebuffi et al., 2020] et Score-CAM [Wang et al., 2020a]. De part son caractère global et peu coûteux en terme d'exécution, Grad-CAM s'est montrée la méthode la plus pertinente dans le cas d'étude, mais il est à noter que Score-CAM permettait d'obtenir des résultats plus précis encore au prix du temps de calcul.

- Des comparaisons avec les réseaux de segmentation ont permis d'établir que le classificateur, malgré qu'il ne soit pas spécialisé proprement dans la tâche de segmentation, montre des résultats particulièrement compétitifs et plus instructifs que ses concurrents non explicables.

- Des comparaisons avec l'application *a posteriori* des différentes CAMs sur un classificateur pré-entraîné ont démontré que les CAMs ne permettent pas d'expliquer proprement les décisions prises par ce classificateur. En effet, les cartes de localisation obtenues étaient à la fois floues et très générales et ne parvenaient pas à cibler les motifs de mouvements de terrain les plus simples. Ceci donne plus de poids encore à la stratégie ECSPLAIN proposée dans cette thèse puisque celle-ci permet au réseau de bien se focaliser sur les zones d'intérêt et de rendre les CAMs exploitables pour l'utilisateur.

- Des comparaisons de combinaisons de CAMs ont également été menées. Elles consistent à entraîner un classificateur avec la stratégie ECSPLAIN selon une CAM donnée et à tester l'explicabilité *a posteriori* de ce classificateur avec les autres CAMs. Parmi les différentes conclusions tirées de cette expérience, il est remarquable que les CAMs restent applicables et bien plus instructives que lorsqu'elles sont appliquées *a posteriori* d'un entraînement standard de classificateur.

Cette méthode est enfin utilisée afin d'enrichir le jeu de données ISSLIDE proposé auparavant. En effet, le classificateur entraîné avec la stratégie ECSPLAIN selon la Grad-CAM a été appliqué sur les interférogrammes présents dans le jeu de données brut d'ISSLIDE. Afin de ne conserver que les détections les plus fiables, une routine de traitements visible en Figure 6.11 a été mise en place. Premièrement, le classificateur est appliqué à l'interférogramme entier à l'aide d'une fenêtre glissante. Une fois les détections obtenues, un moyennage temporel des détections est effectué pour obtenir les zones de mouvements qui se répètent dans plusieurs interférogrammes. Suite à un seuillage et à un filtrage lié à la taille des instabilités identifiées, il est alors possible d'obtenir une carte de segmentation sur toute la surface couverte par les interférogrammes. Une telle routine a permis de trouver plus de 400 potentielles instabilités supplémentaires qui ont été intégrées au jeu de données ISSLIDE avec leur carte de localisation respective.

Conclusion

Contributions proposées...

Comme illustré par cette toute dernière contribution, l'objectif général de la thèse était d'assurer une détection compréhensible et interprétable des instabilités gravitaires tout en profitant de la puissance des algorithmes d'apprentissage automatique et de la vaste couverture des images satellitaires.

Ainsi donc la traduction de modalité SAR-optique a pu être démontrée comme une manière de mieux interpréter les images SAR en rendant disponible une image optique malgré une couverture nuageuse dense. Un point critique qui a motivé l'implémentation de SARDINet a été l'importance de la pertinence des motifs reconstruits et non pas du contraste. En effet, malgré les quelques expériences encourageantes proposées pour améliorer le contraste des images de SARDINet, cela se faisait toujours au prix de la pertinence des motifs reconstruits. Or c'est cette pertinence qui est particulièrement recherchée afin de pouvoir appliquer des algorithmes de détection post-traduction.

Même si cette caractéristique a guidé les choix et décisions prises pour améliorer les performances de SARDINet, les expériences de détection de changements appliquées aux images traduites ont montré des résultats infructueux. Malgré tout, en guidant plus intelligemment le réseau traducteur, notamment avec une tâche de classification (TCL) ou de détection de changements (TCDL) annexe, les performances de traduction ont montré un vrai regain d'intérêt. En effet, même si les premières expériences ne permettent pas d'obtenir de meilleures performances qu'une détection de changements monomodale SAR, elles sont tout de même particulièrement encourageantes puisqu'elles ont permis d'améliorer les performances multimodales initiales.

Enfin, les dernières contributions ont permis l'identification des instabilités gravitaires dès les premiers signes de mouvement. En effet, en calculant des interférogrammes et en mettant au point une stratégie d'identification avec des géomorphologues pour générer ISSLIDE, il a été possible de mieux comprendre ces instabilités pour adapter au mieux les algorithmes à appliquer. C'est ainsi qu'une nouvelle approche, EC-SPLAIN, mêlant détection, explicabilité et localisation a permis de concilier les attentes de compréhensibilité des géosciences avec la puissance des algorithmes d'apprentissage automatique.

... et perspectives futures !

Fort de toutes les expériences, conclusions et résultats obtenus lors des différentes contributions engagées dans cette thèse, de nombreuses voies ont pu être explorées mais de nombreuses autres voies prometteuses ont également été ouvertes.

En effet, du point de vue de la traduction de modalité, les premières expériences se sont concentrées sur la pertinence des résultats et ce avec succès. Les performances obtenues par les modèles entraînés de façon adversaire ont tous montré de sérieuses défaillances au cours de l'entraînement (instabilité) ou dans les reconstructions (artefacts). Ainsi la grande perspective résultant de ces travaux de traduction de modalité consiste à encourager le développement de modèles non adversaires afin d'assurer avant tout la qualité des images traduites avant leur crédibilité. Pour ce faire, plusieurs possibilités sont ouvertes :

- Des changements architecturaux pour mieux transférer les propriétés structurelles des images SAR dans les reconstructions optiques,
- L'utilisation d'informations complémentaires telles que le DEM, la carte des pentes, les masques de distorsions ou encore la composition de la couverture du sol,
- L'intégration intelligente de ces informations complémentaires, soit par fusion de données, soit par conditionnement (*e.g.* à l'aide de conditionnement des couches de normalisation de lots (CBN en anglais) [de Vries et al., 2017, Dumoulin et al., 2017]).

Malgré tout, et comme démontré dans les derniers travaux de traduction de modalité, l'utilisation de réseaux entraînés de façon adversaire reste possible mais à condition soit de les régulariser (*e.g.* avec la TCL), soit de leur confier des tâches plus légères et plus adaptées (*e.g.* le défloutage dans la stratégie Tra-De).

Concernant la détection de changements, les résultats obtenus sur les traductions n'ont pas été fructueux. Mais néanmoins, nous restons convaincus de l'exploitabilité de cette méthode pour mieux interpréter et comprendre les caractéristiques de terrain sous-jacentes. Ainsi plusieurs directions de recherches allant dans ce sens peuvent être considérées :

- Assurer un meilleur guidage de la traduction de modalité surtout lorsqu'elle est appliquée à des données particulièrement difficiles comme celles du jeu de données d'Haïti où la résolution des images est de 10 mètres, la couverture du sol est essentiellement forestière et les glissements de terrain sont très peu visibles sur les images SAR.
- Mettre en place des stratégies d'apprentissage progressif [Bengio et al., 2009] pour assurer en premier lieu un apprentissage fiable sur les échantillons avec peu de difficulté puis complexifier la tâche en ajoutant des images avec des nuages et/ou des distorsions géométriques. De cette façon, l'apprentissage se focalisera avant tout sur la tâche assignée puis intègrera les difficultés dans un second temps.
- Faire une traduction de caractéristiques plutôt qu'une traduction d'images. Cette stratégie basée sur des réseaux par hallucination [Hoffman et al., 2016], à ne pas confondre avec les GANs, permet d'extraire des caractéristiques optiques à partir d'une image radar. De cette façon, la complexité de la tâche de traduction est réduite à l'espace latent d'un encodeur, ce qui permet donc d'alléger la complexité de la tâche ciblée par le réseau.
- Identifier un espace commun. Plutôt que de forcer les images SAR à "devenir" optiques, l'idée de l'apprentissage représentatif [Jain et al., 2021] peut également être exploitée. Celle-ci consiste à projeter les caractéristiques SAR et optiques dans un espace commun qui conviendrait aux deux modalités. De cette façon, la tâche de traduction se fait dans les deux sens et non pas seulement du SAR vers l'optique.

Enfin, la détection d'instabilités lentes est d'autant plus intéressante qu'elle requiert la compréhension du phénomène et l'interaction entre les sciences informatiques et les géosciences. Ainsi pour renforcer ces interactions et en faire profiter les algorithmes d'apprentissage automatique, les premières pistes consistent à intégrer la routine établie pour ISSLIDE dans le modèle :

- Mettre en place des modèles multi-temporel exploitant des séries temporelles d'imagette d'interférogramme avec la même base de temps. Cette stratégie permettrait d'exploiter la répétitivité attendue des mouvements et donc de renforcer la fiabilité des détections.
- Mettre en place des modèles sur plusieurs bases de temps. En utilisant une même date de référence, l'idée principale consisterait à utiliser un interférogramme à 6, 12 et 18 jours pour vérifier la présence du mouvement d'une base de temps à l'autre.
- Intégrer des connaissances physiques à l'entraînement telles que le calcul de l'étendue de la zone, la vitesse de déplacement dans la ligne de visée et celle probable au sol en prenant en compte l'inclinaison de la pente.

In fine, il est important de mentionner que l'objectif de toutes ces propositions est la réalisation d'un algorithme de prévention permettant d'étudier l'évolution du déplacement au cours du temps et alerter en cas de danger de déclenchement imminent.

Finalement, plus que de "simplement détecter des glissements de terrain et phénomènes semblables sur des images satellitaires à l'aide d'intelligence artificielle", les travaux de cette thèse ont consisté à

assurer l'exploitabilité, la compréhensibilité et l'efficacité d'un apprentissage automatique pour que les géosciences, l'informatique et la télédétection collaborent en symbiose.