



THÈSE DE DOCTORAT DE SORBONNE UNIVERSITÉ
Spécialité **Informatique**
École Doctorale Informatique, Télécommunications et Électronique (Paris)

Optically Variable Device Modeling and Verification on Video Captured Documents

Modélisation et vérification de marques optiquement variables sur des captures vidéo de documents

Présentée par
Glen POULIQUEN

Pour obtenir le grade de
DOCTEUR de SORBONNE UNIVERSITÉ

Présentée et soutenue publiquement le 24 mars 2026, devant le jury composé de :

M. Camille Kurtz *Professeur, Université Paris Cité*	Président du jury
Mme. Petra Gomez-Krämer *Maître de conférences, La Rochelle Université*	Rapporteure
M. Sébastien Marcel *Professeur, Université de Lausanne (Unil)*	Rapporteur
M. Oriol Ramos Terrades *Maître de conférences, Universitat Autònoma de Barcelona (UAB)*	Examinateur
M. Thierry Géraud *Professeur, Laboratoire de Recherche de l'EPITA (LRE)*	Directeur de thèse
M. Joseph Chazalon *Maître de conférences, Laboratoire de Recherche de l'EPITA (LRE)*	Co-encadrant de thèse
M. Guillaume Chiron *Docteur, IDnow*	Co-encadrant de thèse
M. Yannis Avrithis *HDR*	Invité
M. Ahmad Montaser Awal *Docteur, IDnow*	Invité

Glen POULIQUEN: *OVD Modeling and Verification on video captured documents*, © 2026

ABSTRACT

This thesis addresses the automated verification of Optically Variable Devices (OVDs) on identity documents captured with commodity smartphones in remote Know Your Customer (KYC) scenarios. OVDs, often referred to as holograms, are key visible security features, as they often overlap and help protect the photo portrait that binds the identity document to its bearer. While current industrial pipelines rely heavily on attack detection (e.g., screen recapture, photocopies), they provide limited guarantees that the expected OVD is actually present and behaves as specified. In this work, we adopt a complementary model-verification perspective: given the document model, we seek to confirm that the portrait-region OVD exhibits valid appearances, sufficient variety, and plausible temporal transitions, under realistic capture constraints (short, low-frame-rate clips, heterogeneous devices, scarce fraud samples).

The contributions are threefold. First, we review existing datasets for identity document analysis and fraud detection, then introduce MIDV-DynAttack, a public, test-only extension of MIDV-Holo with 1,200 additional videos featuring realistic static and dynamic attacks against OVDs on identity documents. We also define an evaluation protocol with identity-disjoint splits and sequence-level metrics, designed to measure generalization to unseen attacks.

Second, we propose two innovative frame-level OVD verification methods: 1) A weakly supervised contrastive approach that learns a generic OVD representation from video-level labels only, trainable without fake data, enabling appearance conformity and coverage checks without frame annotations; 2) After observing that a naive frame classifier exploits dataset biases (e.g., print artifacts) rather than holographic content, we propose HoloVerif, a specific-OVD verifier that combines background subtraction to isolate holographic signal, pseudo-labeling, and a frame classifier focused on OVD behavior. Experiments on MIDV-Holo, MIDV-DynAttack and other datasets show that HoloVerif achieves state-of-the-art performance, especially on dynamic attacks where existing methods fail.

Third, we explore sequence-level modeling to capture transition validity. We investigate two approaches trainable from legitimate videos only: 1) a span-based classifier using synthetically corrupted spans as negative samples; 2) a masked sequence modeling method that uses reconstruction error in an embedding space as an anomaly score.

Overall, this thesis advances the state of the art by introducing the first public benchmark for dynamic attack evaluation, developing methods that achieve strong performance on this challenging task, and laying the groundwork for temporal modeling approaches that address fundamental limitations of frame-level verification.

RÉSUMÉ

Cette thèse porte sur la vérification automatisée des dispositifs optiquement variables (OVD - *Optically Variable Devices*), ou marques optiquement variables (MOV), présents sur les documents d'identité. Ces documents sont capturés par téléphone mobile lors de scénarios de *Know Your Customer* (KYC) à distance. Les OVD constituent un élément de sécurité visuel majeur, car ils recouvrent et protègent souvent la photographie d'identité qui lie le document à son titulaire. Si les solutions industrielles actuelles reposent principalement sur la détection d'attaques connues (recapture d'écran, photocopies), elles n'offrent que des garanties limitées quant à la présence effective de l'OVD attendu et à la conformité de son comportement. Ce travail propose d'y remédier en adoptant une approche de vérification de modèle : étant donné le modèle du document, nous cherchons à confirmer que l'OVD de la zone du portrait présente des apparences valides, une variété suffisante et des transitions plausibles, sous des contraintes d'acquisition réalistes (séquences courtes, hétérogénéité des appareils, rareté des échantillons frauduleux).

Les contributions de cette thèse sont triples. Premièrement, nous dressons un état de l'art des jeux de données existants pour l'analyse de documents d'identité et la détection de fraudes, puis nous introduisons MIDV-DynAttack, une extension publique de MIDV-Holo comprenant 1 200 vidéos supplémentaires. Celles-ci présentent des scénarios réalistes d'attaques statiques et dynamiques visant les OVD. Nous définissons également un protocole d'évaluation rigoureux (métriques au niveau séquence, partitions disjointes par identité) conçu pour mesurer la capacité de généralisation face à des attaques inédites.

Deuxièmement, nous proposons deux méthodes innovantes de vérification d'OVD opérant au niveau image : 1) une approche contrastive faiblement supervisée permettant d'apprendre une représentation générique des OVD à partir d'annotations au niveau vidéo uniquement, sans nécessiter d'échantillons frauduleux, autorisant ainsi des contrôles de conformité et de couverture des apparences sans annotation fine des images ; 2) suite au constat qu'un classifieur naïf exploite les biais du jeu de données (artefacts d'impression) plutôt que le contenu holographique (qui le rend vulnérable aux attaques dynamiques), nous proposons HoloVerif. Cette méthode de vérification d'OVD spécifique combine une soustraction de fond pour isoler le signal holographique, un pseudo-étiquetage des trames vidéo et un classifieur d'images centré sur le comportement de l'OVD. Les expérimentations sur MIDV-Holo, MIDV-DynAttack et d'autres jeux de données montrent que HoloVerif atteint des performances de l'état de l'art, notamment sur les attaques dynamiques où les méthodes existantes échouent.

Troisièmement, nous explorons la modélisation des transitions au niveau de la séquence. Nous étudions deux approches entraînables uniquement à partir de

vidéos légitimes : 1) un classifieur exploitant des segments de vidéo, utilisant des segments corrompus synthétiquement comme échantillons négatifs ; 2) une méthode de modélisation par masquage de séquence qui utilise l'erreur de reconstruction dans un espace de plongement (*embedding*) comme score d'anomalie.

Globalement, cette thèse fait progresser l'état de l'art en introduisant le premier banc d'essai public pour l'évaluation d'attaques dynamiques, en développant des méthodes performantes pour cette tâche complexe, et en posant les fondements d'approches de modélisation temporelle répondant aux limitations de la vérification image par image.

REMERCIEMENTS

Je remercie Petra Gomez-Krämer et Sébastien Marcel d'avoir accepté de relire ce manuscrit et d'en être les rapporteurs.

Je remercie Camille Kurtz d'avoir accepté d'être examinateur et président du jury, ainsi que pour sa participation à mon comité de suivi individuel. Je remercie également Yannis Avrithis pour son suivi durant ces trois années.

Je remercie Thierry Géraud pour la direction de cette thèse.

Je remercie Oriol Ramos Terrades d'avoir accepté d'être examinateur et de m'avoir accueilli lors d'un échange.

Je remercie Joseph Chazalon pour l'encadrement de cette thèse.

Je remercie Ahmad Montaser Awal et Guillaume Chiron pour l'opportunité qui m'a été donnée, d'abord en stage, puis à travers cette thèse CIFRE, ainsi que pour leur encadrement.

Je remercie plus généralement et chaleureusement les équipes de recherche d'IDnow et du LRE, ainsi que mes amis, pour leur soutien durant ces années.

Enfin, je ne saurais remercier suffisamment mes parents et ma sœur pour leur soutien sans faille tout au long de ces trois années.

CONTENTS

ABSTRACT iii
RÉSUMÉ v
REMERCIEMENTS vii
ACRONYMS xvii
1 INTRODUCTION 1
1.1 Know Your Customer (KYC) . . . 1
1.2 Fraudulent Identity Documents . . . 3
1.3 Security Features of Identity Documents . . . 5
1.4 Importance of Optically Variable Devices . . . 6
1.5 Verifying the Authenticity of Optically Variable Devices . . . 7
1.6 Problem Statement and Operational Requirements . . . 9
1.7 Contributions and Manuscript Structure . . . 11
1.8 Scientific Productions . . . 12
1.9 Ethical Considerations . . . 13
2 DATASETS, EVALUATION PROTOCOL AND DATA GENERATION 15
2.1 Existing Datasets of Identity Documents . . . 15
2.1.1 Datasets With Genuine-only Samples . . . 17
2.1.2 Fraud Datasets Featuring Static Templates . . . 17
2.1.3 Fraud Datasets Featuring Dynamic Templates . . . 19
2.2 MIDV-DynAttack, Our Dataset of Fraudulent Identity Documents 20
2.2.1 Content of the Original MIDV-Holo Dataset . . . 20
2.2.2 New Attacks in MIDV-DynAttack . . . 23
2.2.3 Dataset Creation Protocol . . . 26
2.2.4 Proposed Evaluation Protocol . . . 28
2.3 Synthetic Data Generation . . . 30
2.4 Conclusion . . . 30
3 FRAMEWISE (TRANSITIONLESS) OVD REPRESENTATION 33
3.1 Related Work . . . 33
3.1.1 Approaches Checking Static Appearance Conformity . . . 34
3.1.2 Approaches Looking for Holographic Behavior . . . 34
3.1.3 Approaches Checking Both Appearance Conformity and Coverage . . . 35
3.2 Reproduction of the MIDV-Holo Baseline . . . 36
3.3 Binary Frame Classifiers for OVD Verification . . . 37
3.3.1 Naive Frame-wise Binary Classifier (Direct Classifier) . . . 37
3.3.2 Improved Frame-wise Binary Classifier (HoloVerif) . . . 39
3.4 Shaping OVD Latent Space With Contrastive Learning (WSL) . . . 42
3.4.1 Learning Objective . . . 44
3.4.2 Triplet Sample Selection Strategy . . . 44
3.4.3 Augmentations . . . 46

3.4.4 From Frame Representation to Video Qualification 46
3.4.5 Qualitative Analysis of the Learned Representation 47
3.5 Experimental Evaluation and Results 48
3.5.1 Evaluation Protocol 48
3.5.2 General Results 50
3.5.3 Ablation Studies 57
3.6 Conclusion 60
4 SEQUENCE-WISE (TRANSITION-AWARE) OVD REPRESENTATION 63
4.1 Sequence Binary Classifier (HoloVerif-Span) 64
4.1.1 Preprocessing 65
4.1.2 Training 66
4.1.3 Inference and Aggregation for Video-level Decision 67
4.1.4 Preliminary Results and Analysis 68
4.1.5 Discussion and Limitations 71
4.2 Sequence Anomaly Detection via Masked Sequence Modeling (MSM) 72
4.2.1 Related Work 72
4.2.2 Proposed Method 75
4.2.3 Experimental Validation on Synthetic Data 79
4.2.4 Discussion and Limitations 84
4.3 Conclusion 84
5 CONCLUSION 87
5.1 Summary of Contributions 87
5.2 Critical Analysis of Proposed Methods 89
5.2.1 Analysis: Strengths, Weaknesses, and Recommendations . . 89
5.2.2 Recommended Deployment Strategy 91
5.3 Limitations 92
5.4 Perspectives and Future Work 93
5.4.1 Data Resources 93
5.4.2 Pipeline Improvements 94
5.5 Closing Remarks 95

LIST OF FIGURES

Figure 1.1 Online verification pipeline implemented by IDnow (illustration from idnow.io). The user first captures their document (front and back) and then records a video selfie. The system verifies document authenticity, checks against presentation attacks, and ensures the selfie-portrait match before creating the online identity and linking it to a state-issued identity. 3

Figure 1.2 Different security elements implemented on the French identity document (issued from 2021). 5

Figure 1.3 Optically Variable Device (OVD) overlapping the photo portrait of the French identity document (issued from 2021). Various visual states can be observed (on the right) according to the tilt angles (see graph in the bottom left). 7

Figure 1.4 Two complementary directions for OVD model verification. 8

Figure 2.1 In MIDV-Holo dataset, documents are captured in different places involving various backgrounds and lightning conditions (left). Document localizations (i.e. 4 corners) are annotated on all images allowing rectifications (center). Additionally, we propose to define a region of interest containing part of the face and the holograms in charge of securing it (right). Extracted Region of Interest (ROI) from samples labeled as "Originals" (below) contain more or less visible holographic content. *Identities* (names and faces) are synthetic. 19

Figure 2.2 Our proposed dataset MIDV-DynAttack extends the original MIDV-Holo dataset with 1200 new attack videos. MIDV-DynAttack is designed for testing purposes over unseen attacks only, and not for model training or calibration. . . 21

Figure 2.3 Overview of the original MIDV-Holo sequences. 22

Figure 2.4 Overview of the MIDV-DynAttack sequences we released. 24

Figure 2.5 Printed MIDV-Holo templates, stacked by categories, used for the capture of MIDV-DynAttack. 26

Figure 2.6 Capturing process, in various environments, of different MIDV-Holo augmented templates implementing the following attacks: A) Complete Mask, B) Double Sticker, C) Leaf Holo, D) Natural Light - Plastic, E) LED light, F) Star and World, G) Plain Holo. 27

Figure 2.7 Proposed data partitioning over the MIDV-Holo and MIDV-DynAttack datasets following a 64%/16%/20% ratio for Training/Validation/Test. MIDV-Holo "Vanilla" refers to the core dataset released in Koliaskina et al. 2023. Note that specific attack subsets (e.g., Photo replacement, MIDV-DynAttack) are reserved exclusively for testing. 28

Figure 3.1 Overview of the proposed method. The preprocessing step is common to both training and inference phases. The frame classifier is trained using pseudo-labels (Valid/Non-Valid) generated by a generic holographic content detector. The final decision (Legit/Non-Legit) at the video-level is derived by aggregating individual frame classifications. . . 40

Figure 3.2 Overview of the proposed approach, involving: 1) weakly supervised TRAINING with a specific data selection strategy over the trainset; 2) an INFERENCE pipeline extracting optimized features used to compute the final "Legit/Non-Legit" decision based on thresholding pairwise distances. The DECISION panel illustrates the calibration of the threshold over the validation partition. 43

Figure 3.3 Frame sampling strategy for building triplets [A]nchor / [P]ositive / [N]egative from *Legit* and *Non-Legit* videos. Each *Legit* triplet is sampled from a unique *Legit* video (A and P at t, N at $t+1$). Each *Non-Legit* triplet is sampled from two distinct videos of the same identity; A and P belong to a common video, while N is drawn from a different one. All samples undergo random augmentations. 45

Figure 3.4 Integrated Gradients (Sundararajan et al. 2017) visualization of a training sample. Our method effectively directs the network's attention towards the OVD. In contrast, the standard ImageNet-trained model lacks this focused attribution, highlighting the impact of the proposed weakly supervised strategy. 47

Figure 3.5 Visualization of DINOv2 (Oquab et al. 2023) features (first PCA component) on extracted ROIs of genuine French identity samples (issued from 2021) before and after fine-tuning on the MIDV-Holo dataset. The fine-tuned version focuses more accurately on holographic zones (e.g., the RF symbol in the 6th column). Input images were *a posteriori* anonymized for privacy reasons. 48

Figure 3.6 Receiver Operating Characteristic (ROC) curves of selected methods (*Use Case 1 - Specific*) for the MIDV-Holo "Vanilla" (left) and MIDV-DynAttack (right) test sets, following calibration on a separate set. The bottom row highlights zoomed regions corresponding to low False Positive Rate (prioritizing usability) and high True Positive Rate (prioritizing security) regimes. A *True Positive* corresponds to a correctly detected fraud sample. The transparent area around each curve indicates the variance ($\pm\sigma$) observed over 5 folds (and averaged across 6 seeds for stochastic methods). Curves were generated by varying the final decision threshold. 52

Figure 4.1 Overview of the full HoloVerif-Span approach (train and test) using visual samples from the MIDV-Holo dataset available for training. The rectified document frame with its corresponding holographic layer mask shows extractable ROIs and their nature (valid/diff/no OVD), which drives the span corruption policy during training. 65

Figure 4.2 Per-video fraud scores (averaged over all 5-frame windows) across all folds and categories. Each bar corresponds to a category from MIDV-Holo or MIDV-DynAttack (Legits, background ROIs, static-template attacks, swap attacks, and dynamic-template attacks). The horizontal line marks the operating threshold τ (set to 0.5 in the absence of calibration on a validation set with Non-Legit samples); videos with scores above this line are classified as Fraud. Annotations of the form "Above: a/b (p%)" indicate how many videos in each category are flagged as Fraud. Legits cluster well below the threshold (few outliers), while most static/no-OVD and many simple proxy attacks are correctly rejected. Advanced dynamic and swap-based attacks remain only partially detected, illustrating both the benefits and limitations of transition-based verification. 69

Figure 4.3 Our generic MSM approach illustrated. The *Projector* refers to any model capable of extracting holographic features (e.g., pretrained DinoV2, custom WSL). The Seq2seq model refers to any model capable of catching transition behavior and reconstructing sequences from masked elements (e.g., LSTM, Transformer encoder). 76

Figure 4.4 Visualization of the embedding space learned by the projector model using synthesized samples generated from 4 distinct visual references (alpha-blended). On the left, training without the smoothness regularization term leads to a representational collapse where distinct phases are merged into only 3 clusters. On the right, adding the smoothness loss resolves this issue, producing 4 distinct clusters connected by continuous transitions between temporally adjacent samples. 77

Figure 4.5 Visualization of embedding trajectories in the latent space for anomaly detection. Solid lines represent the input sequences, while dashed lines show the model's reconstruction of masked segments (used to compute the anomaly score). We illustrate four cases: (A) **Legit Reference:** The reconstruction adheres closely to the valid trajectory. (B) **Wrong Appearance:** The input features unseen shapes (Visual refs D, E, F), causing high reconstruction error. (C) **Random Transitions:** The input uses valid shapes but violates the learned transition graph (e.g., direct jumps between non-adjacent states). (D) **Sharp Transitions:** The input lacks the smooth interpolation characteristic of genuine hologram transitions. 82

Figure 4.6 Distribution of the Log-MSE reconstruction error across the synthetic test set. Valid samples (green) generally exhibit significantly lower reconstruction errors compared to fraudulent samples (red), which include wrong appearances, invalid transition paths, and sharp discontinuities. The horizontal dashed line represents the decision threshold τ, calibrated at the 99th percentile of the legitimate distribution (corresponding to a fixed False Positive Rate (FPR) of 1%). 83

LIST OF TABLES

Table 2.1 Summary of public identity document datasets related to fraud detection. *Original source: HCT: HandCrafted Templates (fully synthetic or scanned source, with potential digital edition of the fields or photo), Printed HTC: acquired by a capturing device (e.g., smartphone). Alteration nature: Fully digital (i.e., no recapture) VS Recaptures (i.e., physical re-acquisition of the altered document). * Preprint.* 16

Table 3.1 Comparison of ROC Area Under the Curve (AUC) values ("no tracking" mode) between the original MIDV-Holo (Koliaskina et al. 2023) publication and our re-implementation. The consistency of results validates the fidelity of our reproduction. 37

Table 3.2 Experimental results on MIDV-Holo and MIDV-DynAttack datasets across both use cases. F-scores are reported for subsets containing both Legit and Fraud sequences; Recall and Specificity are reported for non-Legit and Legit-only subsets, respectively. Performance metrics are consistent between the two use cases, with the notable exception of the "Dynamic" subset, which was specifically designed to expose the limitations of methods focused purely on the second use case. 50

Table 3.3 Recall performance ($\mu \pm \sigma$) at 5 FPS for each individual fraud type in the MIDV-DynAttack dataset. All methods were calibrated on the MIDV-Holo validation set by maximizing the F-Score. 54

Table 3.4 Generalization performance on additional unseen datasets. 55

Table 3.5 Generalization capabilities across different network architectures. *Starred methods (`MethodName*`) indicate the variants retained in the main results (Table 3.2).* 56

Table 3.6 Ablation study of the HoloVerif and Direct Classifier methods. The table compares the impact of background subtraction (BG subtraction), pseudo-labeling, and data augmentations. *Starred methods (`MethodName*`) indicate the variants retained in the main results (Table 3.2).* 58

Table 3.7 Ablation study of the WSL method. The variants tested include: removal of augmentations (*WSL without augmentation*), training on legitimate samples only (*WSL only Legit*), applying background subtraction (*WSL*) and simply using pretrained weights without fine-tuning (*ImageNet weights*). A variant of the pipeline including a background subtraction preprocessing step is also listed. *Starred methods (`MethodName`*) indicate the variants retained in the main results (Table 3.2).* 59
Table 3.8 Summary of frame-wise OVD verification methods. Each approach addresses different aspects of the verification problem, with complementary strengths and limitations. 62

Table 4.1 Preliminary results for HoloVerif-Span. Metrics aggregate the test results over the 5 folds. 68
Table 4.2 Synthetic dataset split by identity. Each identity yields one legitimate sequence. For test identities, three additional fraudulent sequences are generated-one per attack type (wrong appearance, random transitions, sharp transitions). 81
Table 4.3 Detection performance on synthetic data. The decision threshold is calibrated to fix the False Positive Rate (FPR) at 1% on legitimate samples. We report the Recall (detection rate) for each specific attack type. 81
Table 4.4 Summary of transition-aware OVD verification methods. Each approach addresses different aspects of the verification problem, with complementary strengths and limitations. 86

Table 5.1 Summary of proposed OVD verification methods. "Specific" refers to verifying a known OVD model; "Generic" refers to detecting any holographic content. "Attack samples" indicates whether fraud data is required for training or threshold calibration. 90

ACRONYMS

KYC	Know Your Customer
eIDAS	Electronic Identification And Trust Services
MRZ	Machine Readable Zone
NFC	Near-Field Communication
PRADO	Public Register of Authentic travel and identity Documents Online
ICAO	International Civil Aviation Organization
OVD	Optically Variable Device
OVI	Optically Variable Ink
DOVID	Diffractive Optically Variable Image Device
MLI	Multiple Laser Image
CLI	Changeable Laser Image
AD	Attack Detection
MV	Model Verification
PAD	Presentation Attack Detection
OCR	Optical Character Recognition
ROI	Region of Interest
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
FPR	False Positive Rate
LBP	Local Binary Patterns
HSV	Hue Saturation Value
MSE	Mean Square Error
WSL	Weakly Supervised Learning
CNN	Convolutional Neural Network
LBP	Local Binary Patterns
GAN	Generative Adversarial Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
VAE	Variational Autoencoder
CGI	Computer Generated Imagery
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
HSV	Hue Saturation Value color space

CHAPTER 1

INTRODUCTION

1.1 Know Your Customer (KYC)

Many activities, in both the physical and digital worlds, require high levels of trust to ensure user safety. While purchasing a newspaper requires little knowledge of the buyer, actions such as accessing medical records, initiating a bank transfer, or applying for a loan typically demand strong user authentication. This prevents impersonation, theft, and abusive access to restricted content, while enabling auditing to combat money laundering, terrorist financing, and other illegal activities. The regulatory framework compelling economic actors, such as companies, to verify the identity of their users and clients is commonly referred to as *Know Your Customer (KYC)* regulations, the specific requirements of which vary across jurisdictions.

Although identity is a complex concept in the abstract, state regulations render the notion of *identity* quite straightforward. In practice, and for the purposes of this work, it reduces to the ability of an individual to possess and present an identity document attesting to their unique status. These documents typically assert identity via names, dates and places of birth, unique national numbers, and biometric details. State-issued documents are therefore of critical importance, as they enable the bearer to claim an identity simply by displaying them. A defining feature of these documents is that they provide a link between the physical attributes of the bearer (such as a face picture or biometric data) and factual information (such as a name or date of birth), thereby enabling to bind user accounts to real, state-issued identities.

With the constant increase in online activities and transaction volumes over the last decades, modern systems are replacing traditional in-person identity proofing with smartphone-based capture, offering a secure and convenient alternative for remote verification. Typical application domains include financial services (e.g., banking, gambling, cryptocurrencies, insurance), age-restricted platforms (e.g., social networks), government portals (e.g., tax declarations), and other non-anonymous services. Consequently, remote identity verification has become the foundation of online trust. The central question for digital trust—**who is my customer?**—is answered by aggregating signals that provide evidence linking a user to a state-backed identity. The authenticity of the state-issued identity document

is therefore critical in establishing the link between an individual's real-world identity and their digital account.

Actors in these domains must comply with constantly evolving KYC requirements. In the European Union, the harmonized Electronic Identification And Trust Services (eIDAS) regulation—implemented in 2014 and updated in 2024—defines three levels of assurance indicating how reliably an identity has been verified: 1) **Low**, providing basic confidence in the identity claim; 2) **Substantial**, implying higher confidence and suitable for many financial services onboarding scenarios; and 3) **High**, requiring robust evidence and secure authentication mechanisms.

To assist companies in aligning with diverse local regulations, global KYC providers (e.g., IDnow, Entrust) are often integrated into service onboarding workflows or used as a mandatory validation step to unlock full account privileges. To achieve high levels of trust, some providers choose—or are mandated—to offload specific controls to trained experts for manual verification.

KYC providers typically deliver checks via commodity devices followisng a guided workflow, often composed of two stages from the user's perspective, as illustrated in Figure 1.1. First is *document capture and verification*: document information is captured via the camera stream and the Near-Field Communication (NFC) reader (if available). Visual content is extracted using computer vision techniques such as barcode decoders, Machine Readable Zone (MRZ) parsers, and Optical Character Recognition (OCR). Authenticity verification is then performed based on visual cues and extracted data. Second is *face capture and verification*: the user records a short selfie video to 1) assess liveness using Presentation Attack Detection (PAD) methods ("Is there a real person in front of the camera?"), and 2) compare that person, using face verification techniques, to the photo portrait extracted from the document ("Is this the correct person?"). This effectively binds the physical person to the document.

Information extraction and face verification have been subjects of extensive research in recent years, and mature solutions for these tasks are now available for production systems. However, **authenticity verification** remains a challenging problem, particularly in remote settings involving commodity devices and uncontrolled capture conditions; this **is the focus of this manuscript**. More specifically, we concentrate on the **verification of security features overlapping the portrait region of identity documents**, as these are pivotal in establishing the secure link between the user and the document.

By focusing on this specific aspect of the pipeline, we set aside the problems of information extraction and face verification. We also exclude two other important components that, while critical in production systems, are outside the scope of this work. First, we assume the obtained signals are free from tampering; i.e., they were not injected via software corruption. Software and communication integrity belong to the domain of computer security, which is not our field of expertise. While this limits our consideration of deepfakes, it does not remove the concern entirely. Second, we restrict our study to the isolated analysis of

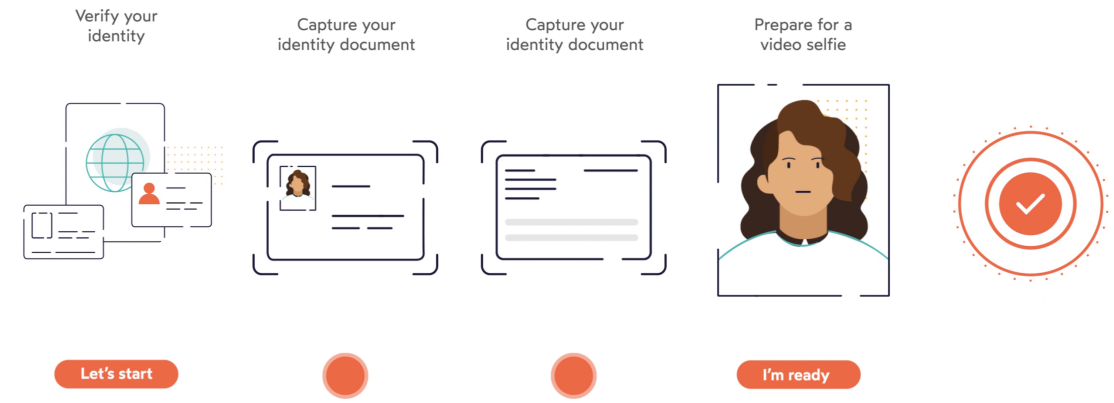

Figure 1.1 – Online verification pipeline implemented by IDnow (illustration from idnow.io). The user first captures their document (front and back) and then records a video selfie. The system verifies document authenticity, checks against presentation attacks, and ensures the selfie-portrait match before creating the online identity and linking it to a state-issued identity.

enrollment sessions (video captures of identity documents via smartphone) and focus on providing confidence indicators. In a full system, these indicators would be aggregated globally, taking into account user activity patterns and other clues to make informed decisions. Such high-level decision-making involves general fraud detection and data analysis techniques that are out of the scope of this work.

To summarize, our general objective is to provide indicators that automatically answer the question "Can this identity document, and the bearer's picture in particular, be trusted?" given a video sample.

1.2 Fraudulent Identity Documents

Fraudsters exploit identity documents to facilitate a wide range of illicit activities, such as: opening bank accounts automatically at scale (e.g., to collect welcome rewards), bypassing age requirements or eligibility checks (e.g., to access restricted content), validating untrackable accounts on cryptocurrency platforms (e.g., for money laundering), and impersonating victims to target specific accounts (e.g., for data theft).

Regardless of whether the victims are states or private entities, these frauds can result in significant financial losses. For example, in 2025, during operation

SIMCARTEL[1], Europol seized 40,000 active SIM cards used for large-scale cyber fraud linked to 50 million fake online accounts, causing losses estimated at €5 million in Austria and Latvia alone.

Practitioners employ taxonomies that categorize fraud into broad families, thereby structuring the verification problem. False documents include *counterfeits* (unauthorized reproductions imitating genuine documents) and *forgeries* (genuine documents that have been unlawfully altered, e.g., via photo or data modification). Additionally, *pseudo* or *fantasy* documents mimic official formats without being officially recognized by any state. Genuine documents can also be subverted: *fraudulently obtained genuine* credentials are issued based on false information or illicit means, while misuse by an impostor involves a "look-alike" presenting another individual's genuine document.

Fraudsters take advantage of the vast and diverse ecosystem of identity documents, which varies significantly across issuing countries and jurisdictions. Since there are numerous official document types (passports, national identity cards, driver's licenses, residence permits, visas)—each featuring multiple versions and occasionally non-homogeneous personalization layers (e.g., due to uncalibrated printers)—verification systems must contend with a wide array of appearances and layout variations. Furthermore, these systems must operate outside the ideal conditions of border control, dealing instead with consumer devices, brief guided sessions, and uncontrolled user environments.

Indeed, the remote nature of online KYC introduces practical constraints that limit what can be verified with certainty. The capture process typically involves short videos, often recorded at low frame rates, with uncontrolled lighting and motion. Unlike in-person verification, there is no opportunity for tactile inspection, and specific illumination sources are rarely available. Consequently, systems must leverage every signal available at capture time—primarily the visual channel, and when present, data-centric cues (e.g., MRZ, barcodes) and device/sensor indicators—to form a reliable judgment.

In this work, we focus on the verification of identity documents specifically to detect *counterfeit* and *forged* documents. We do not consider *pseudo/fantasy* documents, as such cases are routinely filtered out during the early stages of information extraction when templates fail to match known document classes. Finally, this work does not address the issue of *fraudulently obtained genuine* documents; this remains the most challenging form of fraud, as its detection requires contextual information beyond what is available at the visual decision level of our system.

1. Operation SIMCARTEL: https://www.europol.europa.eu/media-press/newsroom/news/cybercrime-service-takedown-7-arrested

1.3 Security Features of Identity Documents

As shown in Figure 1.2, identity documents embed layered security features designed to make inspection straightforward while ensuring that tampering and counterfeiting remain costly and unreliable. This design reflects a defender-attacker dynamic: genuine documents should expose cues that are easy to verify, whereas non-genuine documents should fail under routine scrutiny. In Europe, practitioner guidance and terminology are codified by the Public Register of Authentic travel and identity Documents Online (PRADO) database (European Union 2022), which popularizes the "*Feel, Look, Tilt!*" heuristic—checking for tactile relief, visual detail, and tilt-dependent effects—and provides standardized examples and vocabulary for operators and system designers.

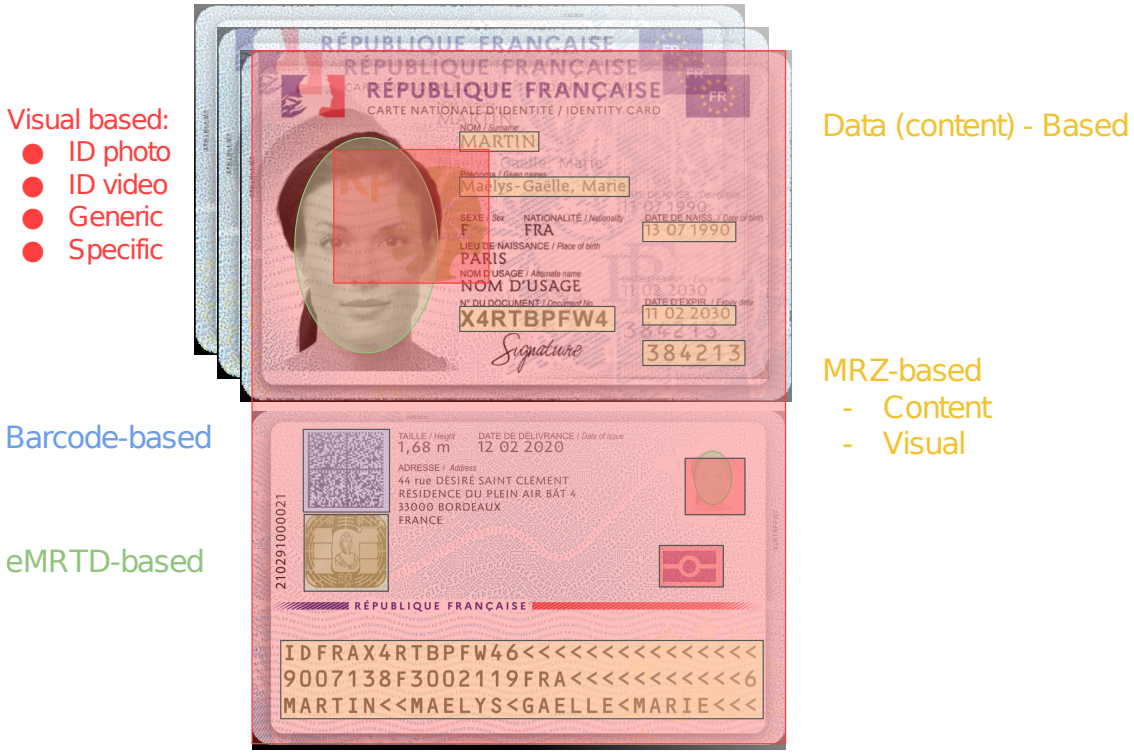

Figure 1.2 – Different security elements implemented on the French identity document (issued from 2021).

Originally designed for in-person inspection by trained operators such as police officers or border guards, many of these features require specialized tools (e.g., magnifiers, UV/IR light sources) or tactile verification. However, in remote settings like smartphone-based KYC, these capabilities are often unavailable. Consequently, the focus shifts to camera-visible cues that can be revealed under ordinary light and slight motion, particularly in and around the portrait where the link between the person and the document is secured.

Furthermore, implementing automated verification in an industrial system without access to detailed specifications of security features presents a significant challenge. Vendors of security devices (e.g., identity document manufacturers) typically keep their designs confidential to prevent fraudsters from reverse-engineering them. As a result, system designers must rely on observable behavior and empirical data rather than detailed technical blueprints. Therefore, we need to develop verification methods that leverage overt, camera-visible features, as covert or restricted features (e.g., tactile elements, UV/IR responses, or advanced NFC content) are generally impractical in remote, smartphone-based scenarios designed by a private KYC provider like IDnow.

1.4 Importance of Optically Variable Devices

Among the various security measures embedded in identity documents, **Optically Variable Devices (OVDs)**—often incorrectly referred to as "holograms"—represent a central visible defense mechanism. Constructed using techniques such as polarized inks and diffraction gratings, they exhibit angle- and illumination-dependent appearances and are commonly positioned over or adjacent to the bearer's portrait. PRADO categorizes these effects under Optically Variable Ink (OVI) and devices (OVD/Diffractive Optically Variable Image Device (DOVID)), as well as related constructs such as *kinegrams* and tilt images (Changeable Laser Image (CLI)/Multiple Laser Image (MLI)) that shift with viewing direction. These overt features are expressly designed for simple "Look/Tilt" checks under ordinary light, enabling reliable human inspection without specialized tools. While this may not fully adhere to the PRADO taxonomy, we will refer to all these technologies as OVDs for simplicity. Figure 1.3 illustrates various visual states (or "appearances") of the OVD overlapping the photo portrait region of the French identity card under visible light at different tilt angles.

These OVDs are of particular interest to us because they constitute the most important feature securing the bearer's picture that can be leveraged in the context of remote identity document verification. However, automating OVD verification using smartphone-captured video presents significant challenges: the signal can be faint or semi-transparent against the face; capture conditions (lighting, motion, glare, surface wear) are uncontrolled; and short, low-fps clips reduce temporal observability. Despite these constraints, OVDs remain the primary camera-visible signals for remote onboarding precisely because they were engineered for straightforward visual checks and serve to secure the bearer's picture. Consequently, in this manuscript, we focus on specific, camera-visible OVDs located in the portrait region. Our aim is to verify that the OVD is effectively present and exhibits most, if not all, of its expected appearances under handheld capture by a remote user.

OVDs are manufactured using proprietary technologies that are kept confidential, with few patents filed publicly. Nevertheless, the manufacturing constraints of



Figure 1.3 – OVD overlapping the photo portrait of the French identity document (issued from 2021). Various visual states can be observed (on the right) according to the tilt angles (see graph in the bottom left).

identity documents appear to leave only two viable options for applying them to the document substrate: either by laser engraving—which enables per-document customization but currently exhibits limited visual behavior—or by the lamination of pre-manufactured thin layers. The latter method produces rich visual behaviors that are strictly identical across all documents of the same series. This consistency suggests it is possible to create a reliable model of these OVDs. This stability is a core characteristic supporting our work: it establishes that it is indeed feasible to know *a priori* which OVD, with its exact visual appearances and possible transitions, is expected for a given document.

1.5 Verifying the Authenticity of Optically Variable Devices

Most operational and academic pipelines emphasize *Attack Detection (AD)*, relying on common PAD algorithms to tackle known attacks (e.g., screen recapture, replay, stream injection) and searching for generic manipulation cues—an effort supported by numerous datasets and benchmarking initiatives. While necessary, AD does not verify that a document conforms to its expected model. In contrast, *Model Verification (MV)* asks whether the specific security feature expected by the document (here, the known OVD) is present and whether its observable behavior matches that model. Since the OVD to be verified is known *a priori* in our setting, the core decision must be driven by MV; AD remains complementary for cases

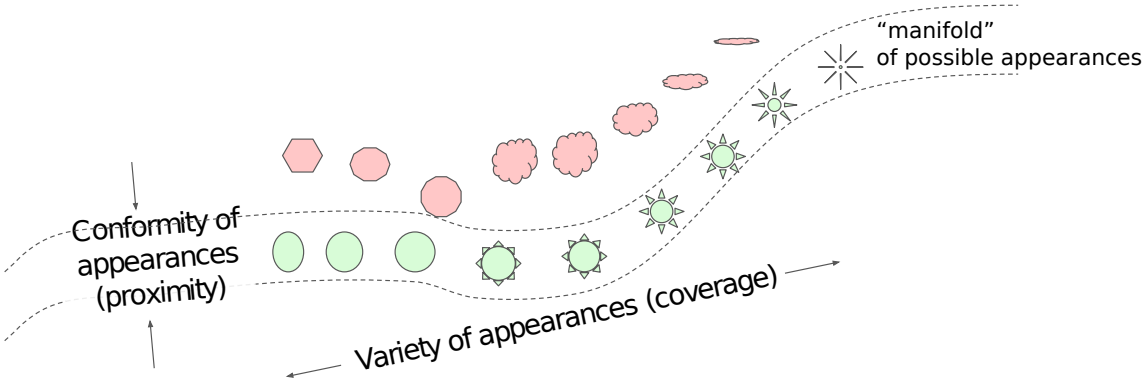

Figure 1.4 – Two complementary directions for OVD model verification.

where specific attack signatures are visible. This dual-paradigm view (authenticity indicators vs fraud indicators) guides our design of a robust identity verification pipeline.

Furthermore, as illustrated in Figure 1.4, we structure OVD MV along two complementary axes:

- **Visual appearance conformity (frame-level).** Each observed frame should exhibit valid, model-specific appearances (shape, color, texture, placement). Relying on this control alone is insufficient, as it can be defeated by static replicas of the expected OVD. Such attacks are described in detail in the next chapter.

- **Visual appearance coverage/variety (sequence-level).** A short video should span a sufficient variety of valid states under small pose and lighting variations. Used alone, this control can be defeated by generic holographic layers that produce dynamic colors without respecting the expected appearance manifold.

When relevant, *visual appearance conformity* can be extended beyond isolated frames to consider short sequences, thereby capturing the dynamic nature of OVD appearances—referred to here as "behavior" or "transitions." Achieving this level of verification is challenging, given the constraints of low frame rates, unpredictable motion, and uncontrolled lighting. However, it holds significant potential for defeating sophisticated attacks, provided that a sufficiently detailed model can be constructed and that the input signals (user-captured videos) are of adequate quality.

Consequently, in contrast to traditional AD approaches, we propose complementing remote identity document verification with MV to enable the detection of unknown manipulations. While this work focuses primarily on MV, we are aware that the development and calibration of our models may occasionally incorporate typical attack signals, blurring the distinction between the two paradigms.

We acknowledge that production systems require the joint application of both paradigms, as MV and AD cover complementary failure modes. For instance, a portrait swap that omits or misplaces an overlapping OVD would be flagged by MV due to non-conforming appearances. Conversely, a highly realistic DeepFake or generated stream might be flagged by AD detecting injection or replay signatures. Similarly, a printed-and-recaptured document might trigger AD (recapture artifacts), whereas MV would detect the absence of the expected OVD behavior. These examples illustrate why a dual-paradigm design is both practical and necessary in industrial applications.

1.6 Problem Statement and Operational Requirements

Problem Statement. The core problem addressed in this manuscript is the automated verification of the authenticity of OVDs overlapping the portrait in remote identity verification scenarios. This involves solving the following tasks:

1. **OVD Model Learning:** Given a set of video samples (genuine, and maybe attacks) of a specific document class, learn a model capturing the expected visual appearances and dynamic behaviors of the OVD overlapping the portrait region.
2. **OVD Model Verification:** Given a short video sequence of a specific document class captured under uncontrolled conditions, determine whether the observed OVD conforms to the learned model in terms of both appearance (isolated frames) or behavior (sequence dynamics)—"Is this the what we expect to see?" and variety over time—"Did we see enough of the different appearances/behaviors?".
3. **Decision Calibration:** Given the outputs of the OVD model verification step, calibrate a final decision policy that meets the operational constraints of industrial identity verification systems (e.g., latency, compute, usability, data availability, generalization).

From an abstract perspective, we can formalize the OVD model verification task as follows: given a short video sequence $V = \{f_1, \ldots, f_T\}$ captured on a consumer device under uncontrolled lighting and motion, and a rectified Region of Interest (ROI) covering the portrait and the adjacent OVD, the objective is to determine a decision $D \in \{\text{Legit}, \text{Attack}\}$ with respect to the known document model.

Operational Constraints. The operational constraints arise from a complex trade-off involving conflicting interests among key stakeholders: **end-users** who may abandon the process if the experience is poor; **service providers** aiming to maximize conversion rates; **regulators** enforcing robust and stringent security

standards; and **industrial operators** managing costs while scaling. These dynamics impose the following specific constraints on the system design:

- *Acquisition.* The capture process involves short video clips, typically recorded at low frame rates under ordinary ambient lighting using handheld consumer devices with heterogeneous optics. Unlike physical inspections, tactile verification is impossible. As a result, inputs must be rectified and cropped to specific regions of interest (ROIs) containing the portrait and the OVD to be usable.
- *Computation and latency.* Resources on both client devices and servers are limited, and end-to-end latency targets are strict to ensure a smooth user experience. This necessitates lightweight backbones, processing focused strictly on critical ROIs, and simple, calibrated decision rules (e.g., single thresholds or concise policies).
- *Usability.* To maintain high conversion rates, the system must require minimal user guidance and cannot rely on specialized illumination or fixed setups. Consequently, algorithms must be robust to unconstrained lighting environments (unknown source positions and intensities) and user execution errors (variable tilt speeds, occasional glare, and imperfect framing).
- *Data availability and annotation.* While genuine samples are abundant, fraudulent samples are scarce, evolving, and expensive to produce. Methods must therefore be capable of learning with limited supervision (state-level or video-level labels rather than frame-level) and reducing dependence on extensive manual annotations.
- *Document model diversity.* The ecosystem includes a vast array of document types and OVD versions. Solutions must be flexible enough to allow specialization for specific high-volume models while maintaining the ability to generalize across broader document families.
- *Generalization.* The system must be robust against *unseen* attack families and shifts in data distribution (e.g., new capture devices). Evaluation protocols must rigorously probe performance in test-only scenarios that extend beyond the curated conditions seen during training.

These requirements delimit the study and guide the design of MV methods and evaluation protocols used throughout the manuscript.

1.7 Contributions and Manuscript Structure

This manuscript is structured around three main axes of contribution: the creation of evaluation resources to overcome the lack of realistic benchmarks, the development of learning-based methods for frame-level verification, and exploratory research into temporal modeling for dynamic verification.

Datasets and Evaluation Protocol. Standard approaches in the literature often rely on datasets that lack realistic dynamic attacks. To address this, we introduce in Chapter 2 MIDV-DynAttack, a test-only dataset comprising 1,200 video sequences. Unlike prior resources, this dataset explicitly targets dynamic attack scenarios, such as manipulating lighting, swapping templates during acquisition, or employing alternate holographic foils. We pair this with a rigorous evaluation protocol that enforces identity separation and reserves specific attack types for testing, actively challenging the generalization capabilities of verification systems.

Learning-based Verification Methods. We investigate data-driven methods to solve the challenge of verifying OVD appearance under the constraints of scarce fraud data. In Chapter 3, we model OVD verification as a frame-level problem, focusing on *appearance conformity*. We propose and compare three distinct approaches:

- A baseline *Direct Classifier* (Direct Classifier), showing that naive supervision can capture useful signals against swap attacks but struggles with generalization.
- A *Weakly Supervised Learning* approach (WSL), which learns discriminative embeddings using only video-level labels. This method captures *appearance coverage* by clustering authentic frames separately from attacks, without requiring frame-wise annotations.
- An *Advanced Classifier* (HoloVerif), which explicitly isolates holographic signals from static content. This method achieves state-of-the-art performance on dynamic attacks by guiding the network to focus on holographic regions.

Exploratory Temporal Modeling. Recognizing that frame-level analysis cannot fully capture the dynamic nature of holograms (i.e., transition validity), we extend our investigation to sequence-level verification in Chapter 4. We explore two directions:

- A *Span Classifier* (HoloVerif-Span), extending our advanced classifier to temporal segments using synthetic temporal corruptions.

- A *Masked Sequence Modeling* approach (MSM), which leverages a two-stage pipeline to detect anomalies in embedding trajectories. This exploratory method demonstrates the extensive potential of unsupervised reconstruction for detecting temporal anomalies without requiring fraud samples during training.

Structure of the Manuscript. The remainder of this thesis is organized as follows:

Chapter 2 surveys existing corpora for document analysis and fraud detection. It details the creation of our MIDV-DynAttack dataset, the specific dynamic attack scenarios it covers, and the evaluation protocol used throughout this work to ensure realistic performance assessment.

Chapter 3 focuses on appearance-centric modeling without explicit temporal reasoning. It details our contributions to frame-level verification, comparing direct supervised classification, weakly supervised contrastive learning, and signal-enhanced specific verification. We analyze the complementary failure modes of these approaches and their dependencies on training data.

Chapter 4 introduces temporal modeling to address transition validity. It presents our exploratory work on segment ("span") classification and masked sequence modeling, discussing the challenges of verifying dynamic behavior on short, low-frame-rate sequences and the potential of synthetic data for training.

Finally, Chapter 5 summarizes the main findings, offers a critical analysis of the proposed methods, and discusses perspectives for future research, emphasizing the need for a layered verification architecture combining multiple paradigms.

1.8 Scientific Productions

The work presented in this manuscript has led to multiple scientific productions, including peer-reviewed conference papers with proceedings and presentations without proceedings. We also detail the public dataset and code released in the attempt to make our research reproducible.

International peer-reviewed publications (with proceedings)

- Glen Pouliquen, Guillaume Chiron, Joseph Chazalon, Thierry Géraud, and Ahmad Montaser Awal. "Weakly Supervised Training for Hologram Verification in Identity Documents". In: *Proceedings of the 18th International Conference on Document Analysis and Recognition (ICDAR) 2024* — **Pouliquen et al. 2024**

- Glen Pouliquen, Joseph Chazalon, Guillaume Chiron, Thierry Géraud, and Ahmad Montaser Awal."Verification of Dynamic Holographic Behavior in

Identity Documents". In: *Proceedings of the 19th International Conference on Document Analysis and Recognition (ICDAR) 2025* — **Pouliquen et al.** 2025a

Presentations without proceedings

The work also led to presentations without proceedings:

- "Identity document layer decomposition," SIFED 2023 — Oral presentation
- "Apprentissage faiblement supervisé pour la vérification des hologrammes dans les documents d'identité," SIFED 2024 — Poster
- "Vérification du comportement dynamique des hologrammes sur les documents d'identité" SIFED 2025 — Oral presentation
- "Modeling Expected Content or Detecting Attacks? Automated Verification of Visual Security Features in Identity Documents" ICDAR 2025 Workshop on Computational Document Forensics (IWCDF) — Keynote

Dataset and code

To support reproducibility and foster further research, we publicly released both the dataset and the code used in our ICDAR publications.

- **Dataset.** The MIDV-DynAttack dataset, introduced in Pouliquen et al. 2025a, is released on Zenodo (Pouliquen et al. 2025b).
- **Code.** All material required to reproduce the results reported in our ICDAR 2024 and ICDAR 2025 papers is publicly available at:
  - `https://github.com/EPITAResearchLab/pouliquen.24.icdar`
  - `https://github.com/EPITAResearchLab/pouliquen.25.icdar`

1.9 Ethical Considerations

Working on identity documents and fraud detection raises important ethical considerations about privacy, fairness, and potential misuse. Throughout this research, we have taken several steps to address these concerns:

- **Data privacy.** All datasets used and figures presented have been either anonymized or based on specimen documents. No personally identifiable information has been collected or shared.

- **Fairness and bias.** Datasets used try to represent diverse demographics and document types to avoid bias. However, we acknowledge that the limited amount of data may not fully capture all variations, and further work is needed to ensure fairness across all user groups.

- **Responsible use.** The methods developed are intended to enhance security and trust in identity verification systems. We designed our proposed approach so they provide calibrated confidence scores, allowing operators to set appropriate thresholds based on their risk tolerance and use cases.

- **Transparency.** We have documented our methods, datasets, and evaluation protocols to promote transparency and reproducibility. This openness allows others to scrutinize and build upon our work.

CHAPTER 2

DATASETS, EVALUATION PROTOCOL AND DATA GENERATION

Chapter abstract

Robust verification of identity documents must withstand unknown and realistic attack scenarios; yet, prior resources lacked the dynamic, sequence-based threats needed to stress-test Optically Variable Device (OVD) verification at scale. We first review existing datasets, and then introduce MIDV-DynAttack, a test-only benchmark of 1,200 videos designed to evaluate generalization to unseen attacks under commodity capture conditions. Complementing this, we define an evaluation protocol with relevant metrics and rigorous train/test splits (across attacks and identities), which will facilitate replication and fair comparison in future research. Together, the dataset and protocol provide a foundation for measuring attack robustness and support subsequent chapters focused on learning and decision strategies. While contributing to improved evaluation of existing approaches against diverse and unseen attacks, MIDV-DynAttack has limitations: it does not introduce new genuine samples—to be classified as *Legit*—but only attack samples; it features only one type of OVD; and while the number of collected samples is substantial, larger volumes would be beneficial. To address these limitations, we introduce at the end of this chapter a work-in-progress approach that leverages synthetic data generation, enabling large-scale production of genuine and attack video samples with realistic appearances, simulating various security devices.

2.1 Existing Datasets of Identity Documents

This section provides an overview of the available datasets for identity document fraud detection, with a specific focus on "holographic" content. Indeed, early identity document datasets, before 2022 were mostly designed for the problematic of content extraction: document localization, template classification, Optical Character Recognition (OCR) and information extraction. Identity document forgeries, whereas being an existing topic of research in the private sector for a while, was address with proper publicly available materials later on. Table 2.1 shows

Table 2.1 – Summary of public identity document datasets related to fraud detection. *Original source: HCT: HandCrafted Templates (fully synthetic or scanned source, with potential digital edition of the fields or photo), Printed HTC: acquired by a capturing device (e.g., smartphone). Alteration nature: Fully digital (i.e., no recapture) VS Recaptures (i.e., physical re-acquisition of the altered document). * Preprint.*

Dataset	No. ids	Doc types	Vids (+Imgs)	Original source (Legits)	Alteration nature (Frauds)	No. Fraud types
Datasets with genuine-only samples						
MIDV-500 Arlazarov et al. 2019	50	50	500 vids	Printed HCT	N/A	
MIDV-2019 Bulatov et al. 2020	50	50	200 vids	MIDV500	N/A	
MIDV-LAIT Chernyshova et al. 2021	180	17	3k imgs	ST-DP-PC	N/A	
MIDV-2020 Bulatov et al. 2022	1k	10	1k+4k	Printed HCT	N/A	
BID Soares et al. 2020	28k	8	0+28k	Digital HCT	N/A	
SpotBID Neves Junior et al. 2024	22k	8	0+22k	Digital fusion	N/A	
DocXPand* Lerouge et al. 2024	25k	15	0+25k	Digital HCT	N/A	
Fraud datasets featuring static templates (e.g., per-image digital alterations, presentation attacks)						
DLC Polevoy et al. 2022	1k	10	1,4k	MIDV2020	Recaptures	3
FMIDV Al-Ghadi et al. 2023	1k	10	28k	MIDV2020	Fully digital	1
KID34k Park et al. 2023	82	2	0+34k	Printed HCT	Recaptures	2
SIDTD Boned et al. 2024	1,2k	10	191+8k	MIDV2020	Fully digital	2
IDNet Guan et al. 2024	41k	20	0+837k	Digital HCT	Fully digital	6
FantasyID Korshunov et al. 2025a	362	13	0+6k	Printed HCT	Fully digital	2
Fraud datasets featuring dynamic templates (e.g., mimic OVD content)						
MIDV-Holo Koliaskina et al. 2023	100	20	700	Printed HCT (+holo layer)	Recaptures	4
MIDV-DynAttack (our) Pouliquen et al. 2025a	100	20	1200	MIDV-Holo	Recaptures + manipulations	10

the progressive shift from genuine identity documents datasets used for content extraction, to more advanced ones based on handcrafted "holographic" templates. This can be structured into three families: 1) datasets containing only genuine samples and designed to assess information extraction performance, 2) datasets containing attack samples without any specific security features embedded in their templates—we call them **static templates**—and designed to assess the performance of edit, replay or presentation attacks[1] detectors, under the general framework of Presentation Attack Detection (PAD) which focuses on the detection of local signal alterations; and 3) datasets containing attack samples on documents secured with OVD-like devices—we call them **dynamic templates**—and designed to also enable Model Verification (MV).

1. *Edit attacks* consist in altering some part of an original document to change some important information (name, date of birth, picture…) and presenting the altered version for verification. *Replay and Presentation attacks* consist in the fraudulent use for verification of a surrogate representation of the original document (like a photocopy or a captured image displayed on screen) that may have been fraudulently obtained.

2.1.1 Datasets With Genuine-only Samples

Non-fraud related datasets of identity documents started to appear from 2018. The first Mobile Identity Document Video (MIDV-500) dataset (Arlazarov et al. 2019) was originally designed to address document analysis tasks (e.g., detection, classification, OCR). It was followed by multiple extensions made by the same team: MIDV-2019 (Bulatov et al. 2020) extending capturing conditions; MIDV-2020 (Bulatov et al. 2022) adding templates personalization; MIVD-Lait (Chernyshova et al. 2021) featuring non-latin alphabets.
The MIDV series is built over handcrafted templates, which are created from scans of identity document samples, anonymized and personalized. Personalized templates are then printed, laminated and captured using various devices (e.g., scanner, handheld smartphones). This process is especially valuable as it tends to reproduce realistic acquisition artifacts (e.g., glare, shadow, motion blur, perspective, occlusion).
Afterward, MIDV-2020 was derived into different altered versions for the need of addressing specific fraud tasks. Depending on the nature of the task, the original MIDV-2020 dataset can therefore be considered as genuine (e.g., when compared to digitally altered samples) or fraudulent (e.g., when compared to advanced handcrafted specimens featuring holographic content as we did in Pouliquen et al. 2024).

Aside from the MIDV series, other datasets of genuine identity documents have been proposed more recently. The BID dataset (Soares et al. 2020) comprises Brazilian identity documents captured under diverse acquisition conditions, whose personal information has been replaced with synthetically generated identities. Its extension SpotBID (Neves Junior et al. 2024) simulates handheld video acquisition by compositing scanned and anonymized BID documents into real backgrounds from the SmartDoc (Burie et al. 2015) video dataset, while synthetically adding controlled spotlight and occlusion effects to emulate challenging illumination. DocXPand (Lerouge et al. 2024) follows a similar spirit, leveraging images collected from real-world production environments in which the original documents are digitally replaced by synthetically generated identities composited into the scene. While providing valuable synthesis pipelines, these datasets do not include any fraudulent samples, and are therefore limited to evaluating the performance of information extraction systems.

2.1.2 Fraud Datasets Featuring Static Templates

Static template fraud refers to alterations made to document templates that do not contain specialized security features. Two categories of alterations can be distinguished.

Fully digital alterations are performed directly on the digital representation of the document and do not involve a physical recapturing step. The general process for such attacks is as follows: genuine documents are captured, their digital versions are altered, and the resulting files are submitted directly for verification. Datasets falling into this category include: FMIDV (Al-Ghadi et al. 2023), which features only crop-and-move (translated copies) of background patches; SIDTD (Boned et al. 2024), which additionally supports inpainting and rewriting (to mask and alter text and signatures); and IDNet (Guan et al. 2024), which further includes text replacement, face morphing, and portrait replacement. More recently, FantasyID (Korshunov et al. 2025a) introduces a benchmark for detecting digital manipulations of ID documents, where cards derived from public sources are printed and re-captured with several devices, then digitally forged by independently altering facial and/or textual regions using modern manipulation methods and this dataset was used in the DeepID challenge (Korshunov et al. 2025b). By assuming that a digital version of the attack can be injected into the verification system, such datasets cover scenarios where users are permitted to upload previously captured images (typically avoided in secure applications) or cases where the acquisition application itself is compromised—a threat model that is generally outside the scope of MV-based document verification.

Recaptures of altered physical signals involve a more complex process where the altered version is recaptured, simulating a scenario where the verification system's acquisition channel is assumed to be secure. The general workflow for such attacks is as follows: similar to *fully digital* attacks, genuine documents are captured and optionally altered digitally (unaltered versions are used to simulate replay attacks). These documents are then either re-materialized (e.g., printed, laminated) or displayed on a screen to be recaptured by the verification device. In many cases, the mere fact that the signal is a recapture is sufficient to classify the data as non-compliant (e.g., due to the presence of moiré patterns) or fraudulent from a KYC perspective. Datasets addressing this problem include DLC (Polevoy et al. 2022) and KID34k (Park et al. 2023), which incorporate captures of 1) reprinted templates that may be altered (e.g., grayscale inputs) and 2) digital versions of the template displayed on a screen. While these datasets are valuable for assessing the detection of known attack signals (such as screen recaptures), they provide limited guarantees regarding generalization to unseen attacks, as the underlying document templates lack specific security features to verify. Consequently, MV methods on these datasets must rely solely on the integrity of the static template features—such as layout, background texture, or font consistency.

Finally, it is worth noting that while these datasets are predominantly video-based, the temporal dimension is often not essential to detect the implemented frauds, as additional views provide little incremental information for static anomalies. At best, the sequential data reinforces the final decision (Legit *vs* Fraud) through aggregation strategies of per-image indicators, such as majority voting.

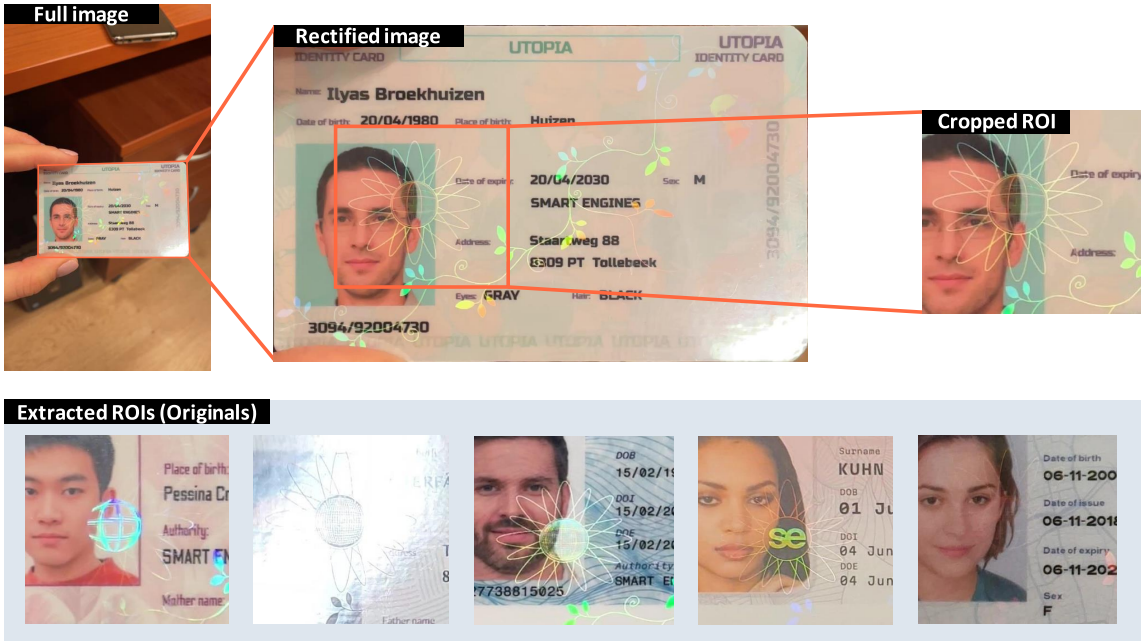

Figure 2.1 – In MIDV-Holo dataset, documents are captured in different places involving various backgrounds and lightning conditions (left). Document localizations (i.e. 4 corners) are annotated on all images allowing rectifications (center). Additionally, we propose to define a region of interest containing part of the face and the holograms in charge of securing it (right). Extracted Region of Interest (ROI) from samples labeled as "Originals" (below) contain more or less visible holographic content. *Identities* (names and faces) are synthetic.

2.1.3 Fraud Datasets Featuring Dynamic Templates

Apart from MIDV-DynAttack—our contribution detailed in Section 2.2—only one publicly available dataset exists as of this writing, MIDV-Holo (Koliaskina et al. 2023), which features identity documents secured with OVD-like devices, as illustrated in Figure 2.1. Indeed, MIDV-Holo features advanced handcrafted dynamic templates, including a holographic layer integrated into the physical support. Less advanced static template copies (e.g., photo overlay, paper photocopy) have also been captured and labeled as frauds. In such dataset, the temporal dimension is essential, as the holographic content exhibits dynamic visual appearances and sequence analysis is required to identify the legitimate holographic behavior. In this context, the source samples used to build the templates before addition of the security layer (e.g., MIDV-2020) can no longer be labeled as Legit (as in the original dataset) but as Fake. As all attack samples have been recaptured, this dataset features a realistic acquisition channel, without relying on the assumption that the acquisition device can be compromised.

As a result, the MIDV-DynAttack dataset we proposed in Pouliquen et al. 2025a (detailed in Section 2.2) aims to complement MIDV-Holo by introducing new attack types that were not originally covered, thereby enabling a more comprehensive evaluation of verification systems' robustness against a wider range of threats, covering non only dynamic template attacks, but also the active use of lighting effects and template *manipulations* during the capture process.

2.2 MIDV-DynAttack, Our Dataset of Fraudulent Identity Documents

This section introduces **MIDV-DynAttack** (published in Pouliquen et al. 2025a), a public dataset that significantly extends the existing MIDV-Holo dataset with 1,200 additional video sequences. MIDV-DynAttack is licensed under the same Creative Commons Attribution-ShareAlike conditions as the original MIDV-Holo dataset. It is designed to challenge automatic OVD verification systems with attacks known to be problematic for existing approaches, yet not covered in MIDV-Holo, as illustrated in Figure 2.2. It is important to note that MIDV-DynAttack is meant to complement the original MIDV-Holo dataset and be used along with it, as illustrated in Figure 2.2. This is why we do not provide any training or validation splits within the MIDV-DynAttack supplementary dataset as *all MIDV-DynAttack is intended to be used as a test set*, with every attack *not seen during training*. Splits are however provided for the MIDV-Holo dataset. Care must be taken to avoid reusing the same identity in more than one of the training, validation or test splits, and our public code cares about proper grouping. Thanks to this protocol, MIDV-DynAttack can be effectively used as a generalization test under unknown attacks.

In the following, we begin by reviewing the content and limitations of the original MIDV-Holo dataset in Section 2.2.1, before detailing the new attacks introduced in MIDV-DynAttack in Section 2.2.2 and the creation protocol in Section 2.2.3. Finally, we summarize the evaluation protocol we propose in Section 2.2.4.

2.2.1 Content of the Original MIDV-Holo Dataset

The original MIDV-Holo dataset (Koliaskina et al. 2023) is built around two fictional "Utopia" document types (passport and ID card) designed to follow common International Civil Aviation Organization (ICAO) Doc 9303 formats (*Uniting Aviation: A United Nations Specialized Agency* 2022). The authors define, for each document type, 10 blank templates (passport: 125×88 mm; ID card: 85.6×54 mm) and personalize each template with 5 synthetic identities (text fields such as name and place of birth, a synthetic portrait, and a matching signature). This results

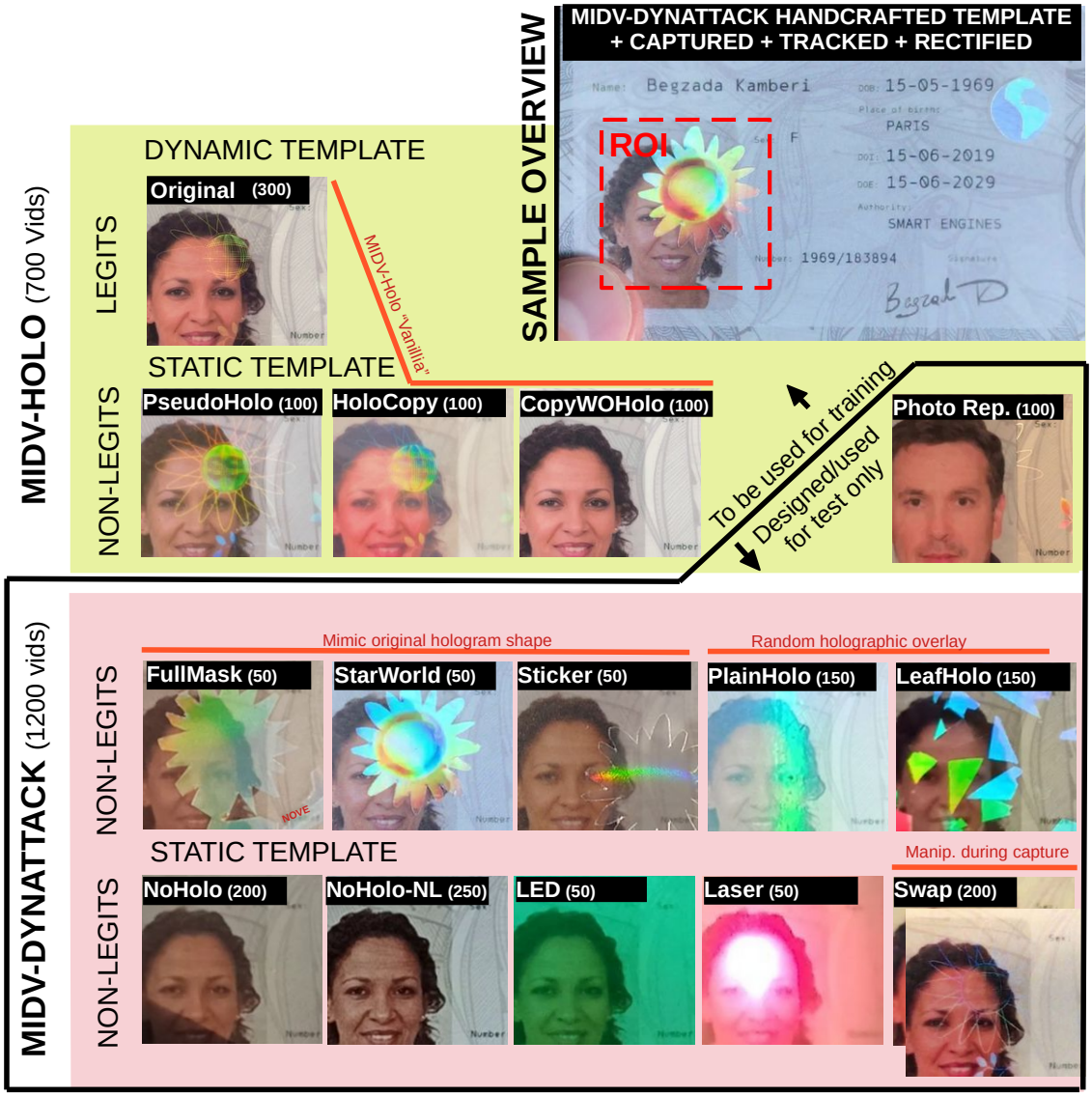

Figure 2.2 – Our proposed dataset MIDV-DynAttack extends the original MIDV-Holo dataset with 1200 new attack videos. MIDV-DynAttack is designed for testing purposes over unseen attacks only, and not for model training or calibration.

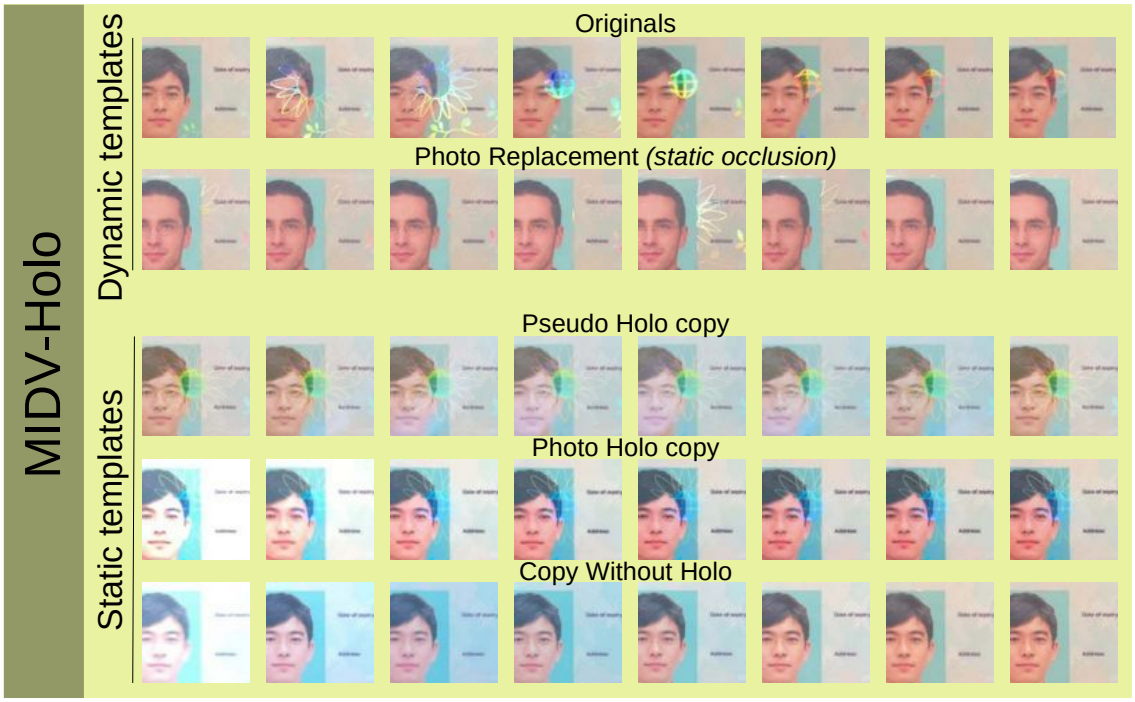

Figure 2.3 – Overview of the original MIDV-Holo sequences.

in 100 distinct document specimens (2 types $\times$10 templates $\times$5 identities). Importantly for OVD verification, the dataset uses a single holographic coating design per document type (one for all passports and one for all ID cards), i.e., the OVD layer is shared within each type.

In addition to "original" (i.e., hologram-bearing) documents, MIDV-Holo includes four threat models meant to approximate practical forgery scenarios, with a particular focus on the portrait area where holographic elements are commonly deployed to deter photo substitution. Specifically, the fraudulent classes include: (i) a copy of the template without hologram ("copy without holo"), (ii) a copy where a hologram pattern is manually drawn at the image level ("pseudo holo copy"), (iii) a printed photo of a scanned original document ("photo holo copy"), and (iv) a photo-replacement attack where the portrait is physically replaced by a printed and cut-out face of a different person attached onto an otherwise original document ("photo replacement").

All sequences are acquired with commodity smartphones (iPhone 12 and Samsung Galaxy S10) across a range of capture conditions designed to promote variability in illumination and specular effects: office lighting with and without deliberate glare, mobile flashlight with and without deliberate glare, and outdoor lighting. Videos are approximately 7 seconds long, recorded in Full HD (1920$\times$1080) or Ultra HD (3840$\times$2160), and then sampled into frames at 5 fps (yielding 33 to 71 frames per clip, with a mean of $\approx$ 43). Overall, the dataset comprises 700 clips: 300 originals and 100 clips for each of the four threat models, with a balanced acqui-

sition protocol across the five lighting conditions. The overview of the sequences is shown in Figure 2.3, focusing on the bearer's picture ROI.

2.2.2 New Attacks in MIDV-DynAttack

We organized the attacks of our MIDV-DynAttack dataset into two main categories: attacks using *static templates*, and the ones using *dynamic templates*. An overview of the new attacks, focusing in the bearer's picture ROI, is shown in Figure 2.4.

Static Template Attacks (750 Videos). The following categories represent traditional forgery techniques using simple printed static templates without any holographic elements. The fraudster relies on lighting effects (e.g., glare, shadows, artificial light sources producing color variations) to create dynamic content in the final video sequence, tricking the system into classifying it as holographic content. These attacks specifically target approaches that estimate variations in hue or saturation in pixel values without verifying the proper visual appearance (shape) of the OVD, or detecting and filtering light spots.

- **Natural Light (NL) - Paper (150 ID cards / 100 passports):** Template without lamination, captures were done under various lighting conditions. Very similar to the "copy without holo" attack from the original MIDV-Holo dataset, but with more aggressive use of light as an attack proxy.
- **Natural Light (NL) - Plastic (200 ID cards):** Template with lamination, half of the captures were done with minimal reflections (100 ID cards), and half were done with deliberate reflections from multiple light sources (100 ID cards).
- **LED-light - Plastic (50 ID cards):** Template with lamination, samples illuminated with multi-color LED-light rings to create different reflective colors.
- **Laser light - Plastic (50 ID cards):** Template with lamination, red laser randomly moving over the document.

We also added a special category that requires more than one template, each sampling a different aspect of the targeted hologram to fake. The fraudster manipulates them in front of the camera so it appears to reproduce the original dynamic holographic content. Such an attack can exhibit several valid visual appearances and aims to defeat approaches that check for multiple valid visual appearances without verifying that actual holographic behavior is present, nor actively prevent improper manipulation during the capture.



Figure 2.4 – Overview of the MIDV-DynAttack sequences we released.

- **Swap - Paper (100 ID cards / 100 passports):** Templates without lamination. There are two different modes: swapping between two printed documents (50 ID / 100 passports) and swapping between three printed documents (50 ID).

Dynamic Template Attacks (450 Videos). The following categories represent advanced forgeries involving handcrafted dynamic templates. In this category, all templates were color printed, modified (e.g., by adding additional holographic layers or elements), and recaptured. These attacks aim to defeat approaches that effectively look for holographic behavior, may check for a variety of visual appearances, but only loosely control the conformity of each visual appearance of the OVD. We separated attacks with generic holographic behavior from those trying to reproduce the actual OVD shape.

- Random holographic overlay (300 videos):
  - **Plain holo (100 ID cards / 50 passports):** Homogeneous holographic material displaying continuous diffraction patterns.
  - **Leaf holo (100 ID cards / 50 passports):** Material exhibiting distinctive polygonal light patterns, shaped as leaves.
- Mimicking the original holograms (150 videos): Targeted forgeries attempting to replicate specific hologram shape and features:
  - **Double sticker (50 ID cards):** Combination of a flower-shaped element cut from laser holographic film and a "genuine" sticker (not overlapping the photo ROI).
  - **Complete mask (50 ID cards):** Rough mask of the hologram applied over the entire printed document.
  - **Star and world (50 ID cards):** Combination of a main flower shape and a world-like hologram pattern.

In terms of complementarity, MIDV-DynAttack introduces attack types that are mostly absent in MIDV-Holo. The only overlap is between MIDV-Holo:CopyNoHolo and MIDV-DynAttack:NL-Plastic. Although MIDV-Holo:Photo-Replacement appears similar to CopyNoHolo, it differs by including a visible hologram fragment on the right side. Additionally, while some MIDV-DynAttack templates exhibit a static holographic appearance (similar to MIDV-Holo:HoloCopy), they were exclusively used for creating swap attacks.

A last point to mention is the lack of new samples that should be accepted as legit—we only produced new attack samples. This is due to the fact we did not



Figure 2.5 – Printed MIDV-Holo templates, stacked by categories, used for the capture of MIDV-DynAttack.

have access to the original holographic layer of MIDV-Holo. Producing a dataset with such content would be a great extension to our work.

2.2.3 Dataset Creation Protocol

Template Creation. MIDV-DynAttack templates were handcrafted using commonly available materials and tools, i.e., MIDV-Holo templates printed, possibly laminated, or supplemented with holographic elements or layers. We aimed to simulate realistic, easily reproducible threat scenarios that could be encountered in industrial workflows, yet causing difficulties for existing verification methods. We did not include any deepfake (or any Computer Generated Imagery (CGI)) samples in MIDV-DynAttack, as we consider this type of attack out of the scope of this work, which focuses on assessing how closely security features of a particular document match a known model, rather than detecting specific attacks (a complementary but different research direction). Indeed, deepfakes require either video stream injection or screen capture, which are covered by other datasets and works (e.g., Polevoy et al. 2022; Park et al. 2023).



Figure 2.6 – Capturing process, in various environments, of different MIDV-Holo augmented templates implementing the following attacks: A) Complete Mask, B) Double Sticker, C) Leaf Holo, D) Natural Light - Plastic, E) LED light, F) Star and World, G) Plain Holo.

Acquisition. All sequences in MIDV-DynAttack were captured at 30 FPS using three smartphone models: an iPhone 7, a Xiaomi Redmi Note 8 Pro, and a Motorola G7. Except for swap attacks (described in Section 2.2.2), which require both hands, devices were naturally handheld to mimic realistic capturing conditions in various environments (e.g., indoor, outdoor). To simulate real-world KYC scenarios, most videos were recorded in naturalistic conditions (except "LED light" and "Laser light"). We deliberately varied capture distances and lighting to reflect uncontrolled authentication environments. Figure 2.5 shows an overview of the printed MIDV-Holo static templates used for the capture of MIDV-DynAttack, and Figure 2.6 shows captures of some templates (i.e., static and dynamic) in various environments (e.g., indoor, outdoor).

Preprocessings (framerate and Registration). To ensure consistency across datasets and mitigate device-specific encoding variations, all video sequences were sampled at 15 FPS. Given that the specific alignment implementation used in the original MIDV-Holo paper is not publicly available, we employed a proprietary industrial-grade rectification pipeline to process both the original MIDV-Holo and the new MIDV-DynAttack sequences. This approach ensures a unified alignment standard and offers improved temporal stability compared to frame-by-frame detection.

The rectification process operates in three stages:

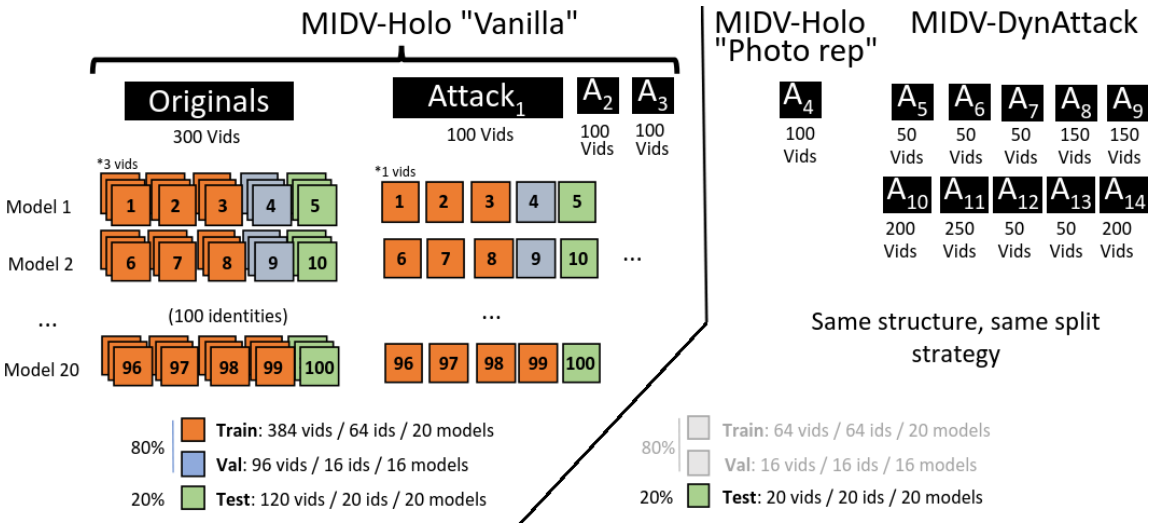

Figure 2.7 – Proposed data partitioning over the MIDV-Holo and MIDV-DynAttack datasets following a 64%/16%/20% ratio for Training/-Validation/Test. MIDV-Holo "Vanilla" refers to the core dataset released in Koliaskina et al. 2023. Note that specific attack subsets (e.g., Photo replacement, MIDV-DynAttack) are reserved exclusively for testing.

1. *Document localization*: The document template is initially matched and localized using a detector based on local image descriptors.
2. *Temporal coherence enhancement*: To handle frames where detection is unreliable (e.g., due to glare or motion blur), a private tracking algorithm is applied. It utilizes a combination of forward and backward tracking strategies to enforce temporal consistency of the document's quadrilateral coordinates.
3. *Trajectory completion*: Any remaining gaps in the sequence where tracking failed are filled via linear interpolation between successfully localized frames.

Following trajectory estimation, a rigid perspective transform is applied to rectify the document in each frame. Finally, the frames are cropped to isolate the specific ROI containing the portrait and the OVD. To foster reproducibility despite the use of a closed-source tracking component, we provide all computed tracking metadata within the dataset release.

2.2.4 Proposed Evaluation Protocol

To foster reproducibility and facilitate fair comparison in future research, we define a rigorous evaluation protocol tailored to the combined usage of MIDV-Holo and our proposed MIDV-DynAttack extension. Integrating the latter as a benchmark requires careful partitioning to prevent data leakage and ensure that

performance metrics reflect generalization capability rather than memorization of specific identities or backgrounds.

Task Definition. The evaluation focuses on the binary classification of video sequences. For a given input sequence showing an identity document, the system must output a decision label: *Legit* (genuine) or *Fraud* (attack). The decision is based on the analysis of the OVD region; while full frames are provided, we standardize the evaluation by extracting the ROI containing the portrait and the holographic layer (using the registration pipeline described previously) to minimize the influence of irrelevant background clutter. The system is permitted to use the entire temporal sequence to inform its decision.

Data Partitioning Strategies. We establish a fixed Train/Validation/Test split designed to assess robustness against unknown identities and unseen attack types. The partitioning logic, illustrated in Figure 2.7, adheres to the following principles:

1. **Identity-disjoint splits:** To ensure valid generalization, we employ a Grouped Stratified K-Fold strategy ($K = 5$). This guarantees that no identity (bearer name/portrait) appears simultaneously in more than one split (training, validation, or test) within a given fold.
2. **Generalization to unseen attacks:** Training and calibration are performed strictly on the "Vanilla" portion of MIDV-Holo, which contains balanced classes of genuine samples and baseline frauds. Testing is conducted on three distinct sets: (i) the held-out "Vanilla" test split (same domain as training); (ii) the "Photo Replacement" subset of MIDV-Holo; and (iii) the entire attacks of MIDV-DynAttack dataset. Crucially, the specialized attacks in (ii) and all samples in (iii) are never seen during training or validation.

Cross-validation and Statistical Reporting. We report performance metrics as the mean $\pm$ standard deviation across the 5 folds. To account for stochasticity in learned methods (e.g., initialization, sampling), we execute 6 runs with different random seeds for each of the 5 folds, resulting in a total of 30 experimental runs. For deterministic baselines, 5 runs (one per fold) are performed. Categorical analysis is performed by first aggregating results within specific fraud types for individual runs, before computing statistical measures across runs.

Model Selection and Calibration. For learning-based approaches, models are trained on the training split, and the best-performing checkpoint is selected based on the validation split. The decision threshold is also calibrated on the validation split to maximize the F_{score}. While learning-based methods update parameters only on the training set, non-learning baselines may calibrate their hyperparameters on the union of the training and validation sets. In all cases, parameters are fixed prior to inference on the test sets.

Metrics. We employ standard binary classification metrics dependent on the composition of the test set:

- F_{score}**:** Reported for datasets containing both *Legit* and *Fraud* samples (i.e., MIDV-Holo Vanilla).
- **Recall (True Positive Rate):** Reported for fraud-only datasets (i.e., MIDV-DynAttack and Photo Replacement subsets), where precision is undefined due to the absence of legitimate samples.
- **Area Under the ROC Curve (AUC):** characterizing performance at low False Positive Rates (FPR). For mixed sets, standard AUC is computed. For fraud-only sets (MIDV-DynAttack), we adopt a two-stage computation: we first determine FPR thresholds from the MIDV-Holo Vanilla ROC curve, then measure the Recall on the attack dataset at those fixed FPR points to integrate the AUC. The final AUC for the combined MIDV-DynAttack and MIDV-Holo dataset is reported as a weighted average based on the volume of video sequences in each attack category.

This protocol, accompanied by the provided split manifests and tracking metadata, ensures end-to-end reproducibility of the results presented in the subsequent chapters.

2.3 Synthetic Data Generation

This section is not publicly available.

2.4 Conclusion

This chapter has defined the data and evaluation backbone used throughout the thesis. Starting from a survey of existing resources, we showed how the MIDV series gradually evolved from generic document analysis datasets to fraud-oriented benchmarks, culminating in MIDV-Holo, the first public dataset explicitly targeting hologram detection. However, these resources only partially meet the requirements identified in the introduction of this manuscript. In particular, they mainly cover static or weakly dynamic attacks and do not fully address the verification of temporal behavior of OVDs under realistic capture conditions.

To move closer to these requirements, we introduced **MIDV-DynAttack**, a test-only extension of MIDV-Holo featuring 1200 additional video sequences that focus on static and dynamic attacks against the OVD region. By design, MIDV-DynAttack complements the original dataset rather than replacing it: it reuses the same handcrafted templates and synthetic identities, but enriches the attack space

with realistic, easily reproducible threat scenarios, including complex dynamic overlays. On top of this data, we specified an evaluation protocol with identity-grouped folds, standardized rectification and ROI extraction.

At the same time, it is important to recognize the limitations of this ecosystem with respect to real-world requirements. First, the underlying MIDV templates are handcrafted and do not correspond to actual national identity documents; they are valuable for research, but remain far from the diversity and complexity of real-world document designs and issuance technologies. Second, our work is constrained by the lack of access to additional genuine OVD layers beyond the one embedded in MIDV-Holo. As noted, we were not able to produce new genuine samples and therefore can only measure recall over unseen attacks, not the precision of generalization on new genuine documents. Finally, security regulations around high-security OVDs used by states make it impractical—and often undesirable—to disclose or replicate their exact optical structures for academic datasets. This imposes an inherent upper bound on how close any public benchmark can get to the true operational landscape.

Within these constraints, the contributions of this chapter position us one step closer to realistic evaluation of OVD verification systems. MIDV-DynAttack and the proposed protocol enable the community to study robustness to a broader range of realistic attacks, with explicit control over identity leakage and temporal sampling. The synthetic/CGI pipeline sketched in Section 2.3 offers a complementary tool to explore counterfactual behaviors on security-relevant ROIs without exposing sensitive content and deal with. In the following chapters, we build on this foundation to develop and analyze learning-based MV methods which, while evaluated on these surrogate datasets, are designed with the more demanding requirements of real-world identity verification in mind. Bridging the remaining gap—by incorporating more diverse genuine samples and closer approximations of state-grade OVDs under appropriate security procedures—remains an important direction for future collaborative work between academia, industry, and issuing authorities.

Chapter 3

FRAMEWISE (TRANSITIONLESS) OVD REPRESENTATION

Chapter abstract

This chapter addresses two distinct use cases for hologram verification: *1) Specific OVD verification*, which involves analyzing the shape, appearance, and behavior of a known hologram model to ensure conformity; and *2) Generic holographic content detection*, a fallback option used when the specific model is unknown or data is insufficient. Robust verification requires representations capable of capturing both visual appearance and dynamic behavior within the hologram's region of interest, while mitigating biases related to the background or identity. We begin by reviewing related work and detailing our public reproduction of the MIDV-Holo baseline. We then introduce three data-driven approaches designed to learn from video-level annotations without requiring expensive frame-wise labels: a naive frame classifier serving as a baseline; an advanced framework (HoloVerif) that leverages background subtraction and pseudo-labels to strictly focus on holographic signals; and a weakly supervised contrastive learning method (WSL) that shapes a discriminative latent space. Extensive evaluations on the MIDV-Holo and MIDV-DynAttack datasets demonstrate that while generic detection methods fail against dynamic forgeries, our specific verification approach achieves state-of-the-art performance and strong cross-dataset generalization. The results highlight the complementarity of these methods and underscore the necessity of verifying appearance conformity to counter sophisticated attacks, paving the way for the temporal analysis explored in the subsequent chapter.

3.1 Related Work

Existing approaches enabling the verification of Optically Variable Device (OVD) content differ primarily in their verification objectives and depth of analysis. Early methods focused on modeling and verifying a set of known, static visual appearances, essentially performing an *appearance conformity* check (i.e., verifying that the displayed shape matches the expected one, using the terminology introduced in Section 1.5). More recent approaches have shifted towards assessing whether

a video Region of Interest (ROI) exhibits generic "holographic" behavior. These methods attempt to circumvent the high cost of frame-wise annotation by substituting the rigorous *appearance coverage* check (i.e., verifying that a sufficient range of expected OVD appearances are displayed) with a weaker proxy based on color statistics. However, very few approaches attempt to simultaneously provide guarantees for both *appearance conformity* and *appearance coverage*. This section reviews the state of the art according to this taxonomy.

3.1.1 Approaches Checking Static Appearance Conformity

A first category of methods attempts to model an OVD using limited training data. For instance, Ay 2022 leveraged adversarial training to generate out-of-distribution samples in an open-set recognition framework. This training strategy was facilitated by the precise cropping of the OVDs and their opaque nature, which prevented background interference through transparency. However, as verification relies on a binary classifier trained to discriminate between in-distribution and out-of-distribution samples, this approach essentially classifies isolated frames. It does not assess the variety of appearances that a genuine OVD should exhibit. Consequently, such an approach remains vulnerable to legitimate-looking static replica attacks, where a counterfeit document displays a fixed holographic appearance.
This limitation regarding appearance coverage is addressed by Chapel et al. 2023 through a classifier based on Local Binary Patterns (LBP). This model is trained on manually annotated frames, where labels correspond to specific visual appearances of the OVD. While promising, this method requires extensive annotation effort and is effectively limited to a checklist of restricted visual appearances, constraining its potential for robust coverage control. Nevertheless, it demonstrates that a "checklist" approach is a viable direction for ensuring appearance conformity. Furthermore, this method was not designed for transparent OVDs, which often exhibit complex patterns mixing holographic signals with non-holographic background information. Like the previous approach, it remains vulnerable to static replica attacks.

3.1.2 Approaches Looking for Holographic Behavior

To overcome the limitations of static appearance modeling and avoid the burden of costly frame-level annotations, several works have proposed focusing on the dynamic behavior of holograms rather than their specific static appearances. Kada et al. 2022 introduced a semantic segmentation approach where documents captured in video are registered, and pixel-level classification is applied to detect holograms based on temporal color statistics. While this approach effectively detects regions containing holographic properties, it lacks the specific appearance

classification necessary to verify the presence of a target OVD model. Consequently, it cannot distinguish between different OVDs and is vulnerable to generic holographic foils that produce dynamic colors without adhering to the expected appearance manifold.
Koliaskina et al. 2023 extended this concept by combining semantic segmentation with a global decision stage based on a variance threshold. They also introduced MIDV-Holo, a dataset containing both genuine and forged holograms for evaluation. Despite parameterization options—specifically regarding the ratio of the area exhibiting significant hue and saturation variance—this approach is inherently designed for generic detection rather than specific discrimination. Therefore, it cannot verify a specific OVD model, as demonstrated in Section 3.5. Similar to the previous method, it remains vulnerable to attacks utilizing generic holographic layers.

3.1.3 Approaches Checking Both Appearance Conformity and Coverage

The challenge of learning an accurate, specific model of an OVD was addressed by Soukup and Huber-Mörk 2017 using dedicated acquisition hardware. Their setup involved a smartphone connected to an LED ring capable of capturing banknote OVD appearances under various controlled illumination angles. For each OVD, a CNN model was trained to authenticate a stack of images acquired via this hardware. While applicable in controlled banking environments, the reliance on specialized tools renders this approach unsuitable for remote Model Verification (MV) scenarios, where users capture videos using personal devices under uncontrolled conditions (varying lighting, motion, and backgrounds). Furthermore, the transparent nature of many identity document OVDs complicates the acquisition of clean reference images, as the background and bearer photo remain visible through the hologram, potentially biasing the learned model. Finally, obtaining a sufficient number of physical samples of a specific OVD model to construct such a dataset is rarely feasible, limiting the practicality of this approach in the context of remote identity verification.

This chapter introduces approaches that model transparent OVDs at the frame level. These methods are designed to learn from video-level annotations alone, utilizing available datasets without requiring dedicated acquisition hardware. To the best of our knowledge, no existing approach proposes to model an OVD using a truly dynamic representation—that is, capturing the instantaneous, differential visual appearance of the hologram at each frame. While this represents an ideal representation of OVD physics, the methods presented in this chapter do not yet address this temporal aspect. This more elaborate dynamic modeling will be explored in the subsequent chapter.

3.2 Reproduction of the MIDV-Holo Baseline

Method Summary	
short name	MIDV-Holo-Baseline
use case(s)	Generic holographic content detection (evaluates approximate appearance coverage; ignores appearance conformity).
overview	Heuristic-based video processing (chromaticity accumulation).
training paradigm	None (parameter calibration only).
attack samples	Required for threshold calibration.

The authors of the original MIDV-Holo publication (Koliaskina et al. 2023) introduced a baseline approach—hereafter referred to as **MIDV-Holo-Baseline**—based on the semantic segmentation of video frames. This method identifies pixels within a holographic area based on their color properties and computes their fill ratio as a proxy to verify the hologram's presence. The system outputs a binary decision: • *negative*: the video clip is deemed to contain the expected hologram and is considered "Legit"; • *positive*: insufficient holographic signal is detected, raising an alert for a potential "Non-Legit".

As no implementation of this baseline was publicly released, we developed an open-source reproduction following the authors' specifications. This re-implementation was motivated by the fact that, to date, this method remains the only OVD verification approach evaluated on a realistic public dataset (MIDV-Holo), making it the primary baseline for our comparative analysis. The source code and evaluation protocols are available at `https://github.com/EPITAResearchLab/pouliquen.24.icdar`.

The algorithm relies on detecting generic dynamic color changes inherent to holograms; it does not explicitly model specific shapes or complex visual behaviors. While its thresholds can be calibrated to maximize verification accuracy for a target hologram model, the method is inherently designed for generic holographic content detection rather than the verification of specific OVD models. It is important to note that the original study performed calibration and testing on the entire MIDV-Holo dataset without a distinct data split. The method relies on three primary parameters: S_{thresh}, h_{thresh}, and T.

Our implementation follows the original description with two minor deviations: 1. input images are resized to 1123×709 pixels; and 2. a minimum temporal buffer of five frames is imposed before a verdict is returned (compared to the theoretical two frames in the original paper). We assessed the fidelity of our reproduction by recalculating the results for the "no tracking" mode presented

Table 3.1 – Comparison of ROC AUC values ("no tracking" mode) between the original MIDV-Holo (Koliaskina et al. 2023) publication and our re-implementation. The consistency of results validates the fidelity of our reproduction.

S_{thresh}	30			40			50		
h_{thresh}	0.01	0.02	0.03	0.01	0.02	0.03	0.01	0.02	0.03
Original MIDV-Holo (Koliaskina et al. 2023)	0.795	0.825	0.832	0.828	0.841	0.832	**0.847**	0.838	0.807
Our re-implementation	0.838	0.846	0.844	0.855	0.844	0.831	**0.857**	0.826	0.790

in the original paper (Koliaskina et al. 2023, Table 1). As detailed in Table 3.1, our implementation achieves nearly identical performance in terms of Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) and identifies the same optimal hyperparameter configuration ($S_{\text{thresh}} = 50$ and $h_{\text{thresh}} = 0.01$) across varying T values.

3.3 Binary Frame Classifiers for OVD Verification

The previous section highlighted that the MIDV-Holo-Baseline approach is inherently generic and lacks the capability to distinguish between different OVDs. To address this limitation, this section introduces two approaches designed to support the verification of specific OVD models without requiring manual frame-level annotations.

3.3.1 Naive Frame-wise Binary Classifier (Direct Classifier)

Method Summary	
short name	Direct Classifier
use case(s)	Specific OVD verification.
overview	Naive supervised binary classifier (frame-level) + confidences aggregated (video level)..
training paradigm	Supervised (video-level labels applied to frames).
attack samples	Required for training and decision calibration.

Regarding the task of **OVD Model Learning** defined in Section 1.6, the most straightforward strategy is to train a binary classifier to determine whether a given frame exhibits a legitimate OVD appearance. As frame-level labels are typically unavailable in standard datasets, we infer them directly from the sequence label. Consequently, every frame extracted from a sequence labeled as *Legit* (e.g.,

containing the target OVD) is treated as legitimate, regardless of whether the holographic content is actually visible. Conversely, all frames from sequences labeled as *Non-Legit* are treated as negative samples. To avoid confusion throughout the manuscript, we distinguish between the two levels of analysis. The classifier operates on individual frames—sub-video elements—and returns a *Validity* result (*Validate* / *Non-Validate*). The final decision at the video level is referred to as *Legitness* (*Legit* / *Non-Legit*). This 2-level analysis will be encountered further on other methods.

During the inference phase (**OVD Model Verification** task), the individual validity confidence scores for each frames are aggregated—for instance, by computing the mean over the sequence. A final binary verdict (Legit / Non-Legit) for the video is then determined by comparing this aggregated score against a calibrated threshold.

This baseline method is referred to as **Direct Classifier** in the experimental analysis presented in Section 3.5. Although relying on coarsely estimated labels, it provides a necessary data-driven baseline against which more advanced methods can be measured. Depending on the composition of the training data, this naive approach can theoretically address two distinct use cases:

- **Specific OVD verification:** by labeling only samples belonging to the target OVD model as *Legit*, while treating all others as *Non-Legit*;
- **Generic holographic content detection:** by labeling sequences exhibiting any valid OVD as *Legit*, and hologram-free sequences as *Non-Legit*. In this scenario, generalization capability is expected to correlate with the diversity of OVDs observed during training.

However, the second use case is considered out of scope for this specific method due to the practical scarcity of data exhibiting sufficient OVD diversity to ensure robust generalization.

A notable characteristic of this naive approach is its potential to capture not only the features of the target OVD but also specific artifacts associated with the attack samples used for training. This effectively blurs the distinction between the Attack Detection (AD) and MV paradigms established earlier, as the model may learn to identify forgery-specific traits rather than focusing exclusively on genuine OVD features. While this dual focus might enhance detection performance on known attacks, it raises concerns regarding overfitting, potentially limiting the model's effectiveness against novel forgeries. The complexity of OVD appearances in real-world video—characterized by variable backgrounds, faint and intermittent holographic signals, and significant noise—may exacerbate this risk of bias.

3.3.2 Improved Frame-wise Binary Classifier (HoloVerif)

Method Summary	
short name	HoloVerif
use case(s)	Specific OVD verification.
overview	Supervised binary classifier (frame-level) + confidences aggregated (video level).
training paradigm	Supervised (using pseudo-labels generated from sequence-level labels and OVD activity detection).
attack samples	Optional for training and calibration, but improve performance.

Building upon the **Direct Classifier** approach, we now introduce a more advanced method (published in Pouliquen et al. 2025a) for learning an OVD model, hereafter referred to as **HoloVerif**. Similar to the **Direct Classifier** method, this approach independently processes each frame of the video sequence—specifically the rectified and pre-cropped OVD ROI—to produce a binary sequence-level verdict (Legit / Non-Legit).

The method consists of three key stages, as illustrated in Figure 3.1. First, a preprocessing step (detailed below) is applied to both training and inference samples to enhance the holographic signal, primarily via background subtraction. Next, a detection model is applied to each video frame to identify instances of genuine holographic behavior. This classifier is trained using pseudo-labels, where frames exhibiting visible holographic behavior from *Legit* sequences are treated as positive samples. Finally, the model outputs undergo a decision stage, where a threshold is applied to the ratio of valid frames within the sequence to determine the final verdict. This decision stage requires calibration to tailor the model's sensitivity to specific application requirements, such as varying tolerances for false rejection.

The primary design goal of this method is to mitigate the risk of overfitting to attack-specific artifacts—a limitation observed in the **Direct Classifier** approach. Furthermore, it represents a first attempt to better capture the complex visual behavior of holograms by utilizing a differential representation of each frame, obtained through background subtraction to isolate instantaneous changes.

Preprocessing to Enhance the Holographic Signal

This preprocessing step is fully deterministic and parameter-free. Given a rectified and pre-cropped OVD ROI from the input sequence, two operations are performed. First, **background subtraction** is applied to remove the static document template—including the bearer's photo, which is specific to each sequence—

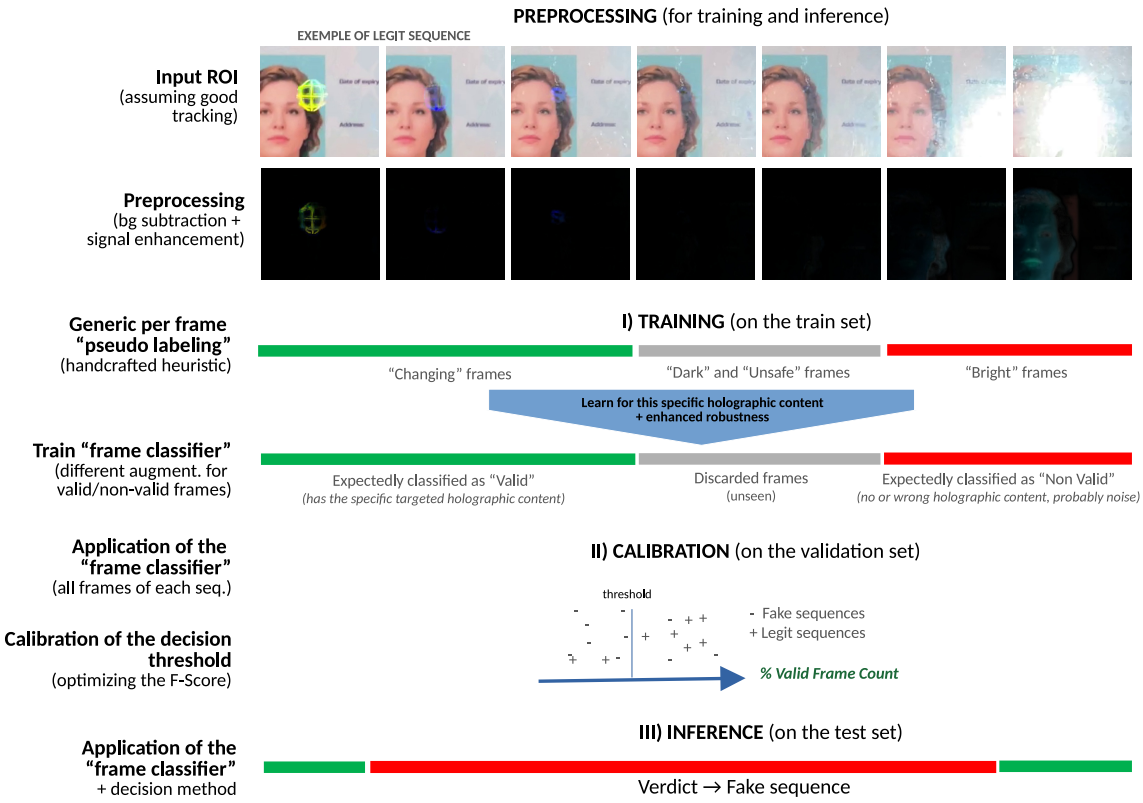

Figure 3.1 – Overview of the proposed method. The preprocessing step is common to both training and inference phases. The frame classifier is trained using pseudo-labels (Valid/Non-Valid) generated by a generic holographic content detector. The final decision (Legit/Non-Legit) at the video-level is derived by aggregating individual frame classifications.

thereby isolating the variable elements. The background image is estimated over the entire sequence using a per-pixel, per-channel median filter and is subtracted from each frame using the absolute difference. The quality of the rectification process (not detailed in this section), which precisely aligns every frame of the sequence, is critical for the efficacy of this subtraction.

Second, an Hue Saturation Value color space (HSV)-based **normalizing filter** N is calculated for each pixel p of each frame I_t as follows, where S and V represent the Saturation and Value channels respectively:

$$N_t(p) = \frac{I_t^{(S)}(p) \cdot I_t^{(V)}(p)}{\max_{t',p'} \left(I_{t'}^{(S)}(p') \cdot I_{t'}^{(V)}(p') \right)}, \tag{3.1}$$

where t' and p' iterate over all timestamps and pixels in the sequence. This normalizing filter N is applied to each background-subtracted frame I to produce the final signal F used in subsequent stages:

$$F_t^{(c)}(p) = I_t^{(c)}(p) \cdot N_t(p), \quad c \in \{R, G, B\}. \tag{3.2}$$

This multiplication acts as a selective filter, attenuating most regions of the image while preserving those characterized by bright, highly saturated colors. Areas with low Value (dark regions) are suppressed as their brightness component approaches zero. Similarly, areas with low Saturation (e.g., white, gray, or black regions) are reduced as their saturation component approaches zero. Consequently, only regions exhibiting high brightness and high color saturation—constituting the optical signature of genuine holograms—remain prominent in the resulting image.

Frame-level OVD Activity Detector

The second step involves a trained classifier designed to distinguish authentic holographic behavior from noise and attacks. This classifier processes a single preprocessed frame and predicts whether it is *Valid* or *Non-Valid* based on the learned holographic model. The training is supervised and relies on two main components: 1. a pseudo-labeling step to generate training targets indicating whether a given frame exhibits holographic behavior; and 2. image augmentations to facilitate the learning of legitimate OVD variations while rejecting noise.

The automated **pseudo-labeling** step addresses the challenge of obtaining detailed frame-level annotations. To generate training data without human intervention, individual frames are labeled as either *Valid* or *Non-Valid* according to the following process:

1. For both *Legit* and *Non-legit* sequences, the absolute luminance ($L_t(p) \in [0, 255]$) of each preprocessed image F_t (after background subtraction) is computed. It is then thresholded to measure the ratio of change $C_t = \frac{|L_t > T|}{|L_t|}$

in each frame, with $T = 5$. Frames are temporarily classified as: *"Bright"* if $C_t > \frac{1}{3}$; *"Unsafe"* if adjacent frames F_{t-1} or F_{t+1} are *Bright*; *"Dark"* if $C_t < 10^{-4}$ (indicating minimal change); or *"Changing"* otherwise. *Unsafe* and *Dark* frames are discarded to prevent training bias.

2. For *Legit sequences*, *Bright* frames are labeled as *Non-Valid* (assuming saturation or glare), while *Changing* frames are labeled as *Valid*.
3. For *Non-legit sequences* (utilizing static template attacks from the original MIDV-Holo dataset), both *Bright* and *Changing* frames are labeled as *Non-Valid*.

Based on this supervised dataset, **image augmentations** are employed to teach the classifier to recognize variations in valid OVD appearances while rejecting out-of-distribution samples, such as noise and imitation attacks. Augmentations are applied to the preprocessed images (i.e., post-background removal and enhancement) and vary depending on the label. For *Valid* frames, mild augmentations are applied to preserve the integrity of the holographic signal. Conversely, for *Non-Valid* frames, stronger augmentations are utilized, as there is no risk of destroying a valid signal.

3.4 Shaping OVD Latent Space With Contrastive Learning (WSL)

Method Summary	
short name	Weakly Supervised Learning (WSL)
use case(s)	Generic holographic content detection (primary), Specific OVD verification (secondary).
overview	Average embedding similarity.
training paradigm	Weakly supervised contrastive learning (uses video-level labels to create triplets).
attack samples	Required for triplet mining (unless trained on Legits only) and decision calibration.

The verification use cases defined in this work—**1) Specific OVD verification** and **2) Generic holographic content detection**—are addressed respectively in Section 3.3.2 and Section 3.2. However, the MIDV-Holo-Baseline (applied to the generic use case) relies on relatively basic heuristics. This motivates the exploration of a modern, data-driven alternative. In this context, this section presents one of the key contributions of this thesis: a method hereafter referred to as **Weakly Supervised Learning (WSL)**, designed to construct a discriminative latent space representing OVD appearances. This latent space is intended to effectively support generic holographic content detection, while also offering a potential

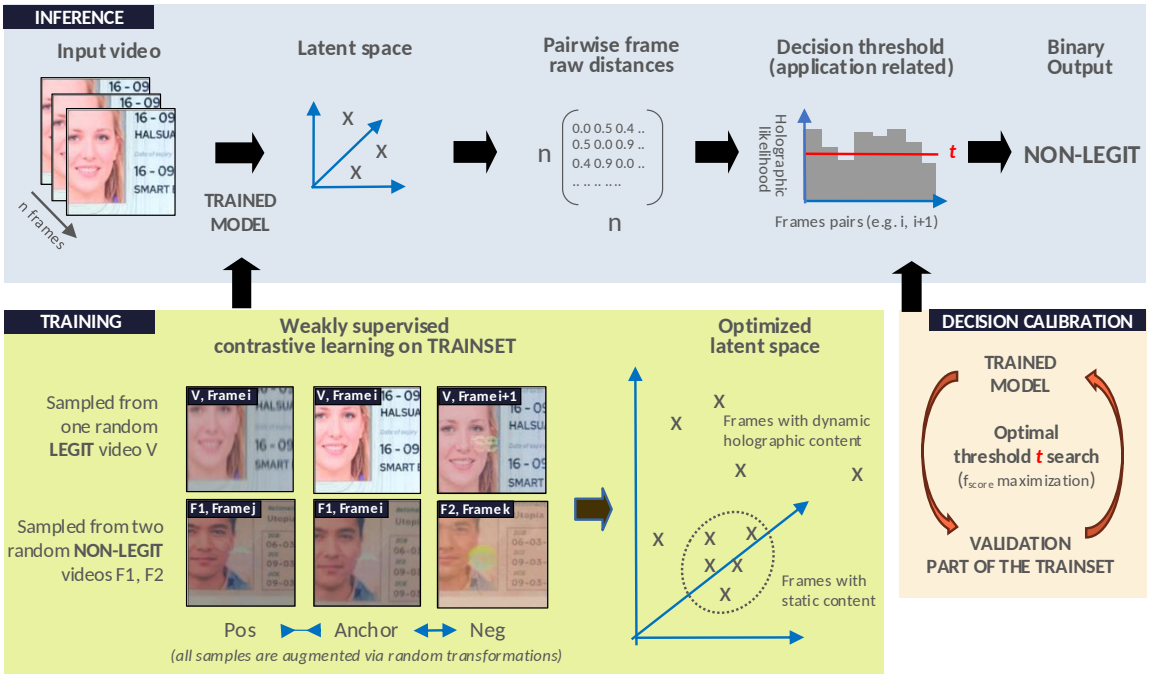

Figure 3.2 – Overview of the proposed approach, involving: 1) weakly supervised TRAINING with a specific data selection strategy over the trainset; 2) an INFERENCE pipeline extracting optimized features used to compute the final "Legit/Non-Legit" decision based on thresholding pairwise distances. The DECISION panel illustrates the calibration of the threshold over the validation partition.

solution for specific OVD verification. A central objective of this method is to learn representations capable of generalizing beyond the specific OVD models encountered during training, thus mitigating data scarcity issues. Similar to the **HoloVerif** described in Section 3.3.2, this approach requires only a single global label ("Legit" or "Non-Legit", following MIDV-Holo (Koliaskina et al. 2023) terminology) per video clip.

Figure 3.2 provides an overview of the method, wherein the OVD representation (extracted features) is learned in a weakly-supervised manner. To this end, we employ a contrastive loss that drives optimization based on intrinsic data properties. This formulation relies on specific assumptions regarding the video clips, such as their ability to capture varied perspectives of the document. The pipeline incorporates various transformation strategies to enhance data diversity. The resulting representation identifies hologram regions effectively and can subsequently be used to assess both appearance consistency and coverage in a final decision stage.

3.4.1 Learning Objective

To avoid the prerequisite of frame-level labels in the training set, we employ a contrastive learning objective that allows model training to be driven by intrinsic data properties (detailed in the subsequent subsection). Specifically, we utilize a triplet loss (Balntas et al. 2016) defined on a minibatch of size N as:

$$\mathcal{L}(a, p, n) = \frac{1}{N} \sum_{i}^{N} l_i(a_i, p_i, n_i), \quad l_i(a_i, p_i, n_i) = \max\left(d(a_i, p_i) - d(a_i, n_i) + m, 0\right) \tag{3.3}$$

where a_i is the projected representation of an *anchor* sample. The objective is to minimize the distance to the representation of a *positive* (similar) sample p_i, while maximizing the distance to the representation of a *negative* (dissimilar) sample n_i. Each representation is computed from augmented inputs to improve robustness. A margin term $m = 1$ is introduced to enforce a minimum separation from negative samples. We employ the Euclidean distance $d(x_i, y_i) = ||x_i - y_i||_2$ and optimize the network using the AdamW optimizer (Loshchilov and Hutter 2019) with default PyTorch parameters.

3.4.2 Triplet Sample Selection Strategy

The selection of samples constituting the triplets is the cornerstone of this approach. It is guided by weak labels provided at the video clip level—specifically, "Legit" or "Non-Legit"—which exhibit distinct temporal properties in the MIDV-Holo dataset. Implicitly, this strategy utilizes video metadata, including the label, video ID, frame timestamp, and (for attack clips) the ID of the falsified identity. For video clips labeled as "Legit", we assume the visual appearance of the hologram shifts throughout the recording. Conversely, for video clips labeled as "Non-Legit", we assume the visual appearance remains static. Crucially, this requires the exclusion of "photo replacement" attacks from the training set, as they exhibit real holographic behavior except at the location of the replaced portrait. This exclusion allows us to evaluate the generalization capabilities of our approach on unseen and challenging attack types. These assumptions underpin the selection process illustrated in Figure 3.3:

- **Legit**: Given that documents in the dataset are captured in motion (sampled at 5 fps), we hypothesize that adjacent frames represent different visual states of the hologram. Thus, frame t and frame $t+1$ are expected to contain divergent holographic appearances. The anchor and positive samples are generated from the same frame t under different augmentations. Frame $t+1$, also augmented, serves as the negative sample.

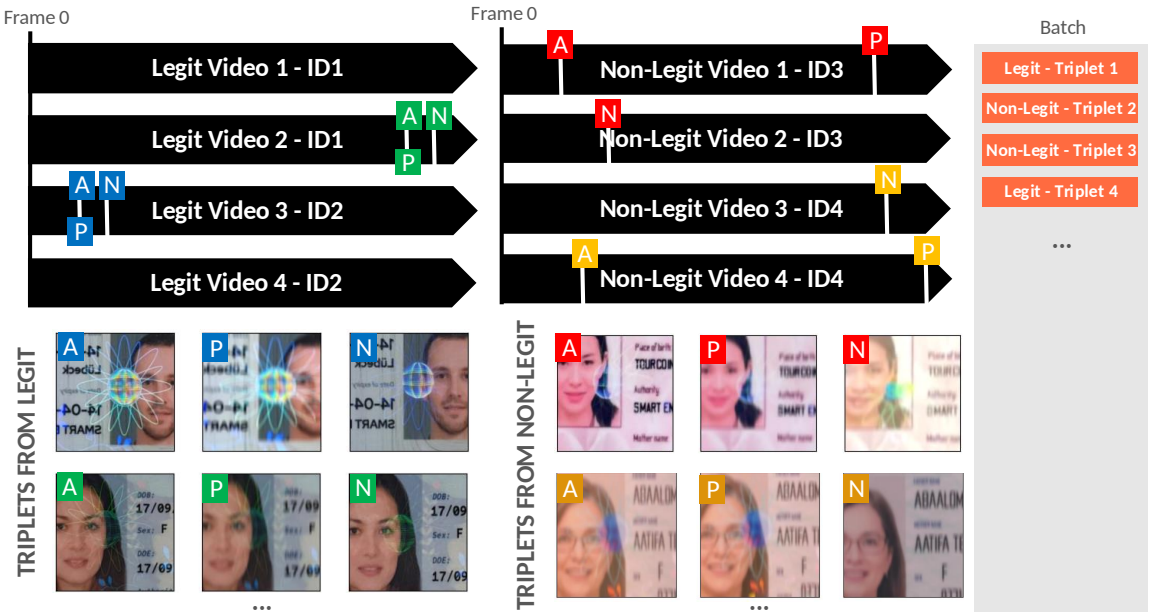

Figure 3.3 – Frame sampling strategy for building triplets [A]nchor / [P]ositive / [N]egative from *Legit* and *Non-Legit* videos. Each *Legit* triplet is sampled from a unique *Legit* video (A and P at t, N at $t + 1$). Each *Non-Legit* triplet is sampled from two distinct videos of the same identity; A and P belong to a common video, while N is drawn from a different one. All samples undergo random augmentations.

- **Non-Legit**: For an Non-Legit video, we rely on the assumption that all frames depict an identical, static holographic state. Therefore, we uniformly select frames from the same video to serve as the anchor and positive samples. For the negative sample, we uniformly select a frame from a substantially different video sharing the same identity (i.e., the same face, but a different recording).

In both scenarios, the anchor, positive, and negative frames depict the same identity (face picture). Consequently, the training objective aims to minimize the distance between embeddings of frames depicting the same instantaneous visual state, while maximizing the distance between frames depicting the same face but different OVD visual appearances (caused by viewing angle changes). Integrating the assumption of continuous viewpoint change—valid in MIDV-Holo—is feasible in real-world scenarios, where document detection stages typically provide pose estimation data.

3.4.3 Augmentations

To diversify anchor, positive, and negative samples (initially resized so that the smaller image dimension is 256 px), we apply the following transformations with the specified probabilities:

- Random cropping of an 80% ROI, resized to 224 px ($p = 1$).
- Gaussian blur ($p = 0.4$, kernel size $k \in [3, 11]$, $\sigma \in [2, 10]$).
- Color jittering ($p = 0.4$, brightness $\in [0.7, 1.3]$, contrast $\in [0.9, 1.1]$, saturation $\in [0.95, 1.05]$).

Images are subsequently normalized using ImageNet mean and standard deviation statistics.

3.4.4 From Frame Representation to Video Qualification

Leveraging the representation produced by the feature extraction network—trained in the weakly-supervised manner described above—we can extract and compare vector embeddings for each frame of a sequence under verification. We subsequently compute the pairwise cosine similarity between these frame representations. By calculating the mean of these pairwise distances, we obtain a single video-level metric indicative of the coverage of expected visual behaviors. A global threshold, pre-calibrated on the validation set, allows for a binary decision: accepting the video as Legit or triggering an alert for a potential Non-Legit.

It is important to note that while this approach does not explicitly classify specific visual appearances, the latent space is shaped such that non-holographic content (which tends to remain constant) is projected to clustered representations, whereas frames containing holographic content (which exhibit variability) are dispersed. Theoretically, this method captures both *Appearance Conformity* and *Appearance Coverage*:

- **Appearance Conformity** is assessed by ensuring that frames containing holographic content map to a distinct region of the latent space, separate from non-holographic approximations. This is enforced by the contrastive objective, which encourages the model to learn discriminative features separating dynamic OVDs from static imitations.
- **Appearance Coverage** is evaluated by measuring the spread of representations for frames containing holographic content. High variability in these embeddings implies a broad range of captured holographic appearances, which is a prerequisite for robust verification. The mean pairwise distance

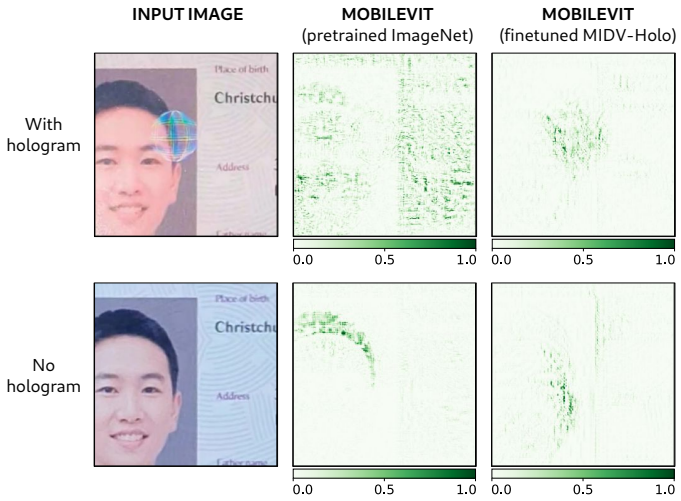

Figure 3.4 – Integrated Gradients (Sundararajan et al. 2017) visualization of a training sample. Our method effectively directs the network's attention towards the OVD. In contrast, the standard ImageNet-trained model lacks this focused attribution, highlighting the impact of the proposed weakly supervised strategy.

metric serves as a proxy for this variability, enabling the system to quantify the extent of holographic content coverage.

3.4.5 Qualitative Analysis of the Learned Representation

Analysis of Learned Features. We employed the Integrated Gradients method (Sundararajan et al. 2017) implemented via the Captum library (Kokhlikyan et al. 2020), to identify the focal elements driving our model's decisions. The results, illustrated in Figure 3.4, demonstrate the efficacy of the weakly supervised training. Specifically, the model ($mobilevit_{xxs}$) trained using our approach assigns significant attribution weight to the hologram area. This stands in sharp contrast to the same architecture trained exclusively on ImageNet, which fails to focus on the holographic signal. This observation underscores the capability of our training strategy to direct the network's attention towards features relevant to OVD verification.

Behavior on Unseen Ovds. We previously alluded to the generalization capability of our method—a property that is non-trivial to verify given that our training data contains only a single genuine OVD model. However, tests on unseen OVD models reveal a qualitative advantage following fine-tuning. Figure 3.5 illustrates this using the first PCA component of DINOv2 (Oquab et al. 2023) features extracted from ROIs of French identity cards, both before and after fine-tuning on the

MIDV-Holo dataset. The fine-tuned version demonstrates significantly improved focus on holographic zones (e.g., the "RF" symbol in the 6th column). DINOv2 was selected as the backbone because this approaches provides powerful feature visualization capabilities.

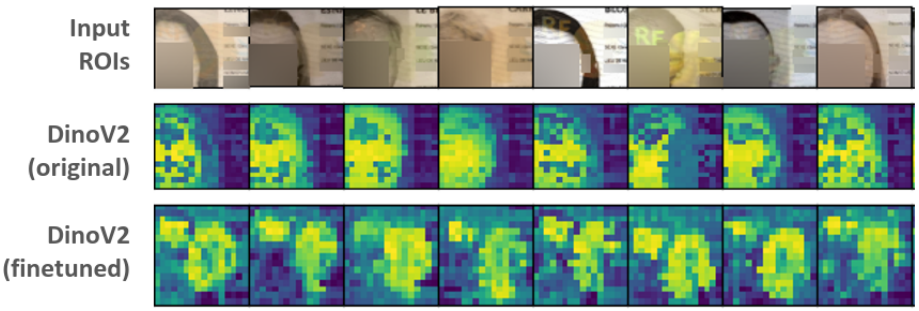

Figure 3.5 – Visualization of DINOv2 (Oquab et al. 2023) features (first PCA component) on extracted ROIs of genuine French identity samples (issued from 2021) before and after fine-tuning on the MIDV-Holo dataset. The fine-tuned version focuses more accurately on holographic zones (e.g., the RF symbol in the 6th column). Input images were *a posteriori* anonymized for privacy reasons.

These results suggest that our method is capable of generating a discriminative latent space that may be sufficiently robust to support specific OVD verification, provided that training data is sufficient.

3.5 Experimental Evaluation and Results

This section presents a comprehensive experimental evaluation of the methods introduced in this chapter. We begin in Section 3.5.1 by summarizing the evaluation protocol—based on the framework established in Section 2.2.4—and listing the comparison methods. Subsequently, Section 3.5.2 details the main experimental results, while Section 3.5.3 reports on the ablation studies conducted to quantify the contribution of individual components within our proposed methods.

3.5.1 Evaluation Protocol

Task. The primary objective of this evaluation is the correct classification of video sequences as either *Legit* or *Non-Legit*, aligning with the **OVD Model Verification** task defined in Section 1.6. This corresponds to the main use case of interest: **Use Case 1: Specific OVD model verification**. Unless otherwise noted, the results reported in this section focus on this specific verification task. An

initial experiment assessing the viability of the fallback scenario—**Use Case 2: Generic holographic content verification**—is provided separately.

Data Splitting Strategy. We adhere to the train/validation/test splits defined in Section 2.2.4, ensuring strict identity separation between partitions. We also remind that MIDV-Holo (train + val splits) is used for training / calibrating all methods, while MIDV-DynAttack (full set) is only used for testing, and that only one OVD model is present in MIDV-Holo.

Training Objective. All methods are calibrated to minimize classification errors at the sequence level (Legit vs Non-Legit), measured using F-score for dataset parts including both classes, Recall for Non-Legit-only subsets, and Specificity for Legit-only subsets—in the specific case of the evaluation of **Use case 2: Generic holographic content verification**.

Metrics and Reporting. In all the tables, metrics are provided for different datasets and reported as mean ± standard deviation over 5-fold cross-validation. For methods including randomness, results are averaged across 6 different random seeds.

Preprocessing. All methods receive as input the same preprocessed OVD ROIs, extracted and rectified using the protocol detailed in Section 2.2.4. The frames were sampled at 5 FPS.

Compared Methods. The following methods are evaluated:

1. **MIDV-Holo-Baseline** (Koliaskina et al. 2023): A baseline method that analyzes pixel-wise chromaticity statistics accumulated over the video stream. We utilize our public reproduction of this method, as detailed in Section 3.2.
2. **Direct Classifier** (Ours): A naive binary frame classifier that can leverage any categorizable signal (including print artifacts or dataset biases) to distinguish between genuine and attack sequences. It serves as a data-driven baseline that does not explicitly model holographic behavior.
3. **HoloVerif** (Ours): The proposed method detailed in Section 3.3.2, which employs background subtraction preprocessing and pseudo-labeling to train a supervised frame classifier focused specifically on holographic content. It is designed to enhance robustness against dynamic template attacks.
4. **WSL** (Ours): A weakly supervised learning approach that utilizes contrastive learning and video-level labels to optimize a feature extraction and classification pipeline. This method aims to learn a discriminative latent space without requiring frame-level annotations.

Table 3.2 – Experimental results on MIDV-Holo and MIDV-DynAttack datasets across both use cases. F-scores are reported for subsets containing both Legit and Fraud sequences; Recall and Specificity are reported for non-Legit and Legit-only subsets, respectively. Performance metrics are consistent between the two use cases, with the notable exception of the "Dynamic" subset, which was specifically designed to expose the limitations of methods focused purely on the second use case.

	MIDV-Holo (test split)			**MIDV-DynAttack** (full)			**Both**
	Vanilla 120 mixed vids*		**Photo Rep.** 20 vids	**Static** 110 vids	**Static-swap** 40 vids	**Dynamic** 90 vids	**Mix** 380 vids
Verification methods ↓	**F-score**	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**[1] **/Specificity**[2]	**AUC**
Main Use Case 1 (specific): Dynamic subset considered as non-Legit (metric: Recall[1]).							
MIDV-Holo-Baseline	85 ± 3	85 ± 12	38 ± 21	72 ± 3	20 ± 8	3 ± 3	57 ± 1
WSL	95 ± 3	96 ± 5	93 ± 8	80 ± 10	15 ± 9	3 ± 2	83 ± 3
Direct Classifier	90 ± 4	89 ± 7	69 ± 18	80 ± 15	89 ± 10	52 ± 20	91 ± 3
HoloVerif	96 ± 2	96 ± 5	59 ± 18	93 ± 5	27 ± 12	61 ± 14	93 ± 2
Fallback Use Case 2 (generic): Dynamic subset considered as Legit (metric: Specificity[2]).							
MIDV-Holo-Baseline	85 ± 3	85 ± 12	38 ± 21	72 ± 3	19 ± 8	97 ± 3^2	-
WSL	95 ± 3	96 ± 5	93 ± 8	80 ± 10	15 ± 9	97 ± 2^2	-

3.5.2 General Results

Suitability for the Defined Use Cases

The experimental results for both *Use Case 1: Specific OVD model verification* and *Use Case 2: Generic holographic content verification* are summarized in Table 3.2. *Use Case 1* remains our primary focus, as it verifies specific appearance coverage, unlike *Use Case 2*, which serves as a less robust fallback option.

Since the MIDV-Holo-Baseline and WSL methods were originally designed with *Use Case 2* (generic detection) in mind, Table 3.2 highlights their performance for both scopes. For *Use Case 2*, the labels of the MIDV-DynAttack *Dynamic* subset must be interpreted differently:

- **Use Case 1 (Specific):** The labels are set to *non-Legit*. This subset contains holographic content that differs from the genuine MIDV-Holo model on which the methods were trained and calibrated. This represents the most restrictive scenario (i.e., verifying a specific document model).
- **Use Case 2 (Generic):** The labels are set to *Legit*. Since the dataset contains sequences with forged yet visually realistic holographic content, a generic detector is expected to validate them. In this scenario, only strictly non-holographic content (e.g., static templates or swap-based frauds) is considered *non-Legit*.

Consequently, the reported metrics in Table 3.2 are identical between the two use cases, except for the "Dynamic" column. This subset was specifically crafted to demonstrate the limitations of MIDV-Holo-Baseline and WSL in verifying specific appearance coverage. Although WSL was conceived with the potential to address *Use Case 1*, its current decision stage does not fully exploit the discriminative latent space. This limitation is explored further in Chapter 4. Nevertheless, WSL proves valuable in the context of the generic fallback (*Use Case 2*), as it clearly outperforms the previously established MIDV-Holo-Baseline baseline.

Because *Use Case 1* (Specific OVD verification) remains our primary research objective, the aggregated AUC metric—which we aim to maximize—provided in Table 3.2 is computed exclusively for this case. Similarly, all subsequent metrics, tables, and figures in this section focus on this specific verification task.

Figure 3.6 provides complementary insights regarding the optimal decision boundaries for the test sets, effectively illustrating the trade-off between security and usability offered by each method. The stark contrast between the left and right sets of ROC curves highlights that MIDV-DynAttack presents a significantly more challenging evaluation environment than the original MIDV-Holo dataset.

Our HoloVerif approach, which deliberately focuses on holographic content, achieves leading performance on MIDV-Holo as well as on unseen static and dynamic attacks from MIDV-DynAttack. Notably, data augmentations substantially improve performance on dynamic attacks, increasing recall from 32% to 61%. While other methods may achieve marginally superior performance on the static attacks of the original MIDV-Holo dataset, they lack this robustness to dyamic variations.

Two specific attack types prove partially challenging for nearly all methods:

1. **Photo Replacement:** Performance is low for all methods except WSL. This is likely because traces of the genuine hologram remain visible outside the area of the replaced face picture. This vulnerability suggests that focusing attention exclusively on critical elements (such as the portrait) is essential. The relative success of WSL here warrants further investigation into why its feature extraction remains robust in this scenario.
2. **Document Swap:** This attack is particularly effective against appearance-based methods. The MIDV-Holo and WSL approaches are deceived by the presence of multiple saturated patterns that satisfy their variance checks. Similarly, HoloVerif fails due to its background estimation strategy; it estimates the backgound frame based on one of the displayed templates, while the others produce a difference image that still exhibits an acceptable holographic appearance.

Combating the Document Swap vulnerability may require detecting the swap itself (e.g., via specific attack detection) or incorporating broader contextual signals (as the Direct Classifier does). However, a more promising direction involves improving hologram modeling through temporal analysis—i.e., learning instan-

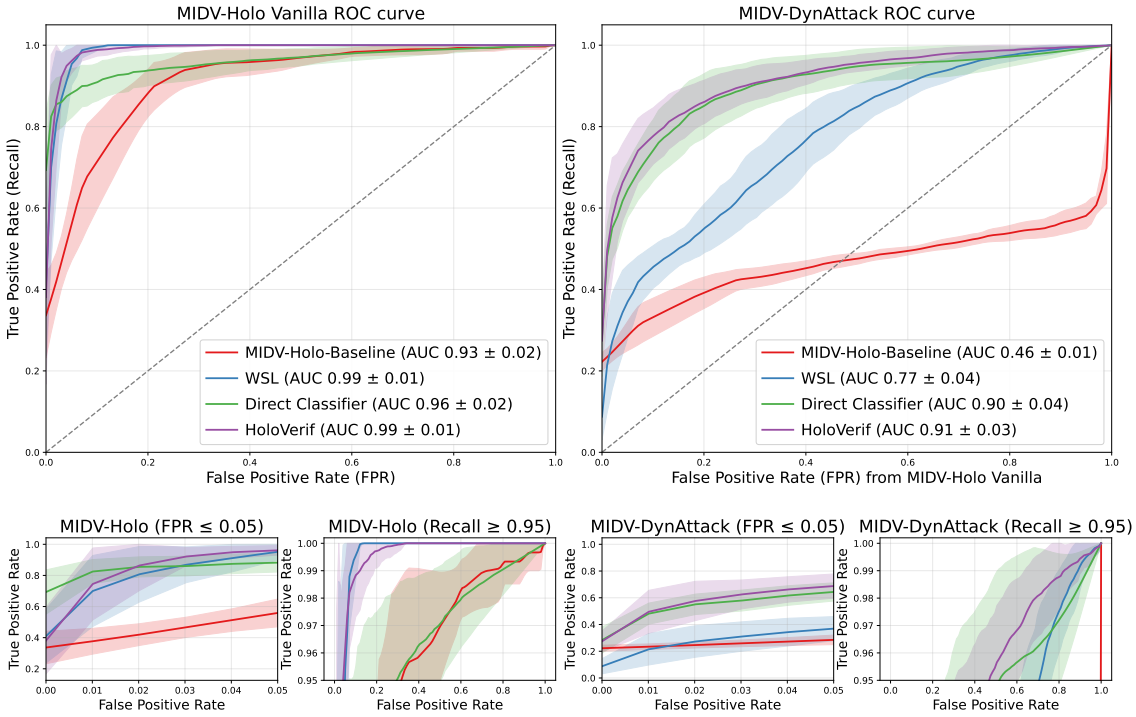

Figure 3.6 – ROC curves of selected methods (*Use Case 1 - Specific*) for the MIDV-Holo "Vanilla" (left) and MIDV-DynAttack (right) test sets, following calibration on a separate set. The bottom row highlights zoomed regions corresponding to low False Positive Rate (prioritizing usability) and high True Positive Rate (prioritizing security) regimes. A *True Positive* corresponds to a correctly detected fraud sample. The transparent area around each curve indicates the variance ($\pm\sigma$) observed over 5 folds (and averaged across 6 seeds for stochastic methods). Curves were generated by varying the final decision threshold.

taneous color transitions from the ordered sequence rather than processing the video as an unordered set of frames.

Key Takeaways. Overall, both the Direct Classifier and the proposed HoloVerif achieve superior aggregated AUC scores (> 90). However, they do not detect the same types of attacks. The Direct Classifier excels at detecting static frauds and captures some dynamic ones, likely by leveraging subtle image signal differences between legitimate and illegitimate samples learned during training. This suggests it could serve as a complementary approach to HoloVerif, which demonstrates leading performance on dynamic attacks.

Per-attack Performance Breakdown

For the remainder of this analysis, we exclusively consider *Use Case 1* (Specific OVD verification). Table 3.3 details the performance of all presented methods on the MIDV-DynAttack dataset, broken down by attack type.

Conventional static attacks (Plastic, Paper) are reasonably handled by all methods. However, Laser and LED light attacks—where an external light source creates saturated patches on a static printout—reveal important limitations even on static frauds that these methods are designed to detect: MIDV-Holo-Baseline achieves 0% recall on both, as these artificial highlights trigger the same chromatic variance the baseline uses as a proxy for holographic behavior. WSL also fails on LED light (1%), confirming its reliance on similar statistics. In contrast, HoloVerif achieves 100% recall on both, demonstrating that learned representations can distinguish genuine holographic patterns from lighting artifacts.

Swap attacks expose the fundamental limitation of background-subtraction-based methods. When multiple templates are alternated, the estimated background converges to one of them, causing the difference images of remaining templates to exhibit seemingly valid holographic signals. Consequently, HoloVerif (18–53%), MIDV-Holo-Baseline (26–0%), and WSL (19–2%) all fail to reliably detect this attack. Direct Classifier is the only method achieving robust detection (93–76%), likely by exploiting dataset biases or identity-specific signals that remain inconsistent across alternating templates.

Dynamic attacks confirm that detecting some holographic behavior is insufficient for robust verification. MIDV-Holo-Baseline and WSL fail almost entirely across all dynamic attack types (0–6% recall). HoloVerif achieves the strongest performance (33–91%), demonstrating that background subtraction combined with targeted augmentations helps the model learn appearance conformity rather than mere holographic presence. The Leaf holo attack remains surprisingly challenging for all methods (1–37%), suggesting visual characteristics that current frame-level representations cannot reliably distinguish from the genuine OVD.

Table 3.3 – Recall performance ($\mu \pm \sigma$) at 5 FPS for each individual fraud type in the MIDV-DynAttack dataset. All methods were calibrated on the MIDV-Holo validation set by maximizing the F-Score.

		Methods (Use Case 1 - Specific)			
	MIDV-DynAttack subsets ↓	MIDV-Holo-Baseline **Recall**	WSL **Recall**	Direct Classifier **Recall**	HoloVerif **Recall**
Static	Plastic (low reflects)	100 ± 0	100 ± 0	92 ± 15	92 ± 8
	Plastic (with reflects)	91 ± 4	80 ± 18	72 ± 21	88 ± 10
	Paper	83 ± 7	90 ± 10	76 ± 16	92 ± 6
	Laser light	0 ± 0	63 ± 33	76 ± 34	100 ± 0
	LED light	0 ± 0	1 ± 3	89 ± 16	100 ± 0
	Swap (two)	26 ± 10	19 ± 12	93 ± 10	18 ± 9
	Swap (three)	0 ± 0	2 ± 4	76 ± 21	53 ± 27
Dynamic	Plain holo	6 ± 6	3 ± 3	74 ± 19	91 ± 8
	Leaf holo	1 ± 2	2 ± 3	37 ± 21	33 ± 20
	Double Sticker	0 ± 0	4 ± 9	36 ± 30	70 ± 20
	Complete mask	6 ± 13	6 ± 6	57 ± 28	55 ± 19
	Star and world	0 ± 0	0 ± 0	42 ± 26	52 ± 22

Key Takeaway: Complementary Failure Modes. No single method succeeds across all attack types. Direct Classifier handles swap attacks effectively but shows inconsistent performance on dynamic frauds; on the other hand HoloVerif exhibits the opposite pattern. This complementarity reinforces that production systems should combine Model Verification with Attack Detection.

Table 3.4 – Generalization performance on additional unseen datasets.

Use Case 1 (Specific)				
Methods ↓	**MIDV-Holo Vanilla** (120 vids) F-score	**MIDV-2020** (1k vids) Recall	**Internal no-OVD** (191 vids) Recall	**Internal OVD** (453 vids) Recall
MIDV-Holo-Baseline	85 ± 3	78 ± 5	80 ± 3	32 ± 3
WSL	95 ± 3	91 ± 7	88 ± 10	34 ± 14
Direct Classifier	90 ± 4	64 ± 19	77 ± 15	61 ± 19
HoloVerif	96 ± 2	96 ± 3	94 ± 4	85 ± 8

Robustness to Unseen Templates and OVD Models

Beyond MIDV-DynAttack, we assess the transferability of each approach to additional datasets that differ in document design, capture conditions, and fraud patterns. The first additional dataset is MIDV-2020 (Bulatov et al. 2022), which contains videos of identity documents without OVDs, with different templates and acquisition setups compared to MIDV-Holo. The second dataset is an internal collection of real identity document videos with or without OVDs, captured using a different device and lighting conditions than those in MIDV-Holo. Both datasets were not used during training or calibration, and should be classified as *Non-Legit* as none of them contain the genuine OVD model present in MIDV-Holo. The cross-dataset generalization results presented in Table 3.4 reveal a clear hierarchy. Methods designed primarily as hologram *detectors* (the MIDV-Holo baseline and WSL) maintain strong F-scores and AUC on the full MIDV-Holo collection, but their recall drops markedly on the MIDV-2020 and external datasets. This confirms that, although effective within their training distribution, their decision functions are tightly coupled to the specific chromatic statistics and acquisition artifacts of the original benchmark. Meanwhile, the Direct Classifier, capable of exploiting any categorizable signal in the images, improves recall on these two unseen datasets compared to WSL. This suggests it captures more generic biases (e.g., print and recapture artifacts), yet it still fails to provide uniformly high recall across all domains.

In contrast, HoloVerif consistently demonstrates the best cross-dataset robustness. It matches or exceeds the strongest competitors on the union of all MIDV-Holo splits, while substantially outperforming them on MIDV-2020 and the external dataset in terms of recall (Table 3.4). This behavior validates its design: by aggressively removing identity- and layout-specific content through background subtraction, and by relying on pseudo-labels derived from generic hologram-detection cues, HoloVerif is forced to learn a representation more tightly aligned with true holographic behavior. Consequently, it is less sensitive to irrelevant domain shifts. Practically, this implies that a single calibration on MIDV-Holo can be reused with relatively minor degradation when the system is deployed on different document families and capture pipelines.

Table 3.5 – Generalization capabilities across different network architectures. *Starred methods (`MethodName`*) indicate the variants retained in the main results (Table 3.2).*

Verification methods ↓	**MIDV-Holo** (test split)			**MIDV-DynAttack** (full)			**Both**
	Vanilla 120 mixed vids		**Photo Rep.** 20 vids	**Static** 110 vids	**Static-swap** 40 vids	**Dynamic** 90 vids	**Mix** 380 vids
	F-score	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**	**AUC**
Use Case 1 (Specific): Dynamic subset considered as non-Legit (metric: Recall).							
WSL mobilenet	89 ± 4	91 ± 10	93 ± 8	70 ± 10	19 ± 10	2 ± 2	69 ± 2
WSL mobilevit*	95 ± 3	96 ± 5	93 ± 8	80 ± 10	15 ± 9	3 ± 2	83 ± 3
WSL resnet	95 ± 2	96 ± 5	93 ± 6	71 ± 10	17 ± 11	3 ± 2	79 ± 3
WSL dinov3	96 ± 2	97 ± 5	91 ± 10	71 ± 16	14 ± 15	1 ± 1	79 ± 5
Direct Classifier mobilenet	80 ± 4	79 ± 10	69 ± 15	76 ± 13	76 ± 19	55 ± 23	84 ± 4
Direct Classifier mobilevit*	90 ± 4	89 ± 7	69 ± 18	80 ± 15	89 ± 10	52 ± 20	91 ± 3
Direct Classifier resnet18	87 ± 3	83 ± 6	70 ± 14	76 ± 12	82 ± 11	54 ± 22	91 ± 3
Direct Classifier dinov3	95 ± 3	94 ± 5	68 ± 21	92 ± 7	87 ± 15	40 ± 16	91 ± 3
HoloVerif mobilenet	92 ± 3	93 ± 6	70 ± 17	92 ± 6	25 ± 12	59 ± 14	90 ± 2
HoloVerif mobilevit*	96 ± 2	96 ± 5	59 ± 18	93 ± 5	27 ± 12	61 ± 14	93 ± 2
HoloVerif resnet18	94 ± 4	95 ± 7	67 ± 22	93 ± 6	25 ± 13	63 ± 18	92 ± 2
HoloVerif dinov3	95 ± 2	96 ± 4	61 ± 22	97 ± 3	34 ± 15	74 ± 14	95 ± 1

Impact of Backbone Architecture

Finally, we examine the impact of backbone architecture on the three main approaches (WSL, Direct Classifier, and HoloVerif). To ensure practical relevance for mobile deployment, we selected the smallest available variant of each architecture family from the `timm` library (Wightman 2019): MobileNetV3-Small (1.6M parameters, Howard et al. 2019), MobileViT-XXS (1.3M parameters, Mehta and Rastegari 2021), ResNet-18 (11.7M parameters, He et al. 2016), and ViT-Small (21M parameters, Siméoni et al. 2025). All backbones were initialized with ImageNet (Deng et al. 2009) pretrained weights, with the exception of ViT-Small which leverages DINOv3 self-supervised pretraining (Siméoni et al. 2025). However, all WSL variants degrade similarly on the MIDV-DynAttack dynamic attacks and on unseen datasets, regardless of model capacity. This indicates that the bottleneck lies in the learning paradigm (video-level contrastive training effectively detecting "any" hologram) rather than the expressive power of a particular backbone. For the Direct Classifier and HoloVerif, increasing backbone complexity (e.g., moving from MobileViT to ResNet-18 or a ViT-based encoder) yields incremental but not transformative gains. Detection of static and some dynamic frauds improves slightly, and aggregated AUC increases by a few points, yet the qualitative ranking between methods remains unchanged (Table 3.5). Significantly, even the smallest models tested are sufficient for HoloVerif to achieve high AUC on MIDV-Holo and strong recall on MIDV-DynAttack. This suggests that once the pipeline is properly structured around hologram-focused preprocessing and supervision, substantial improvements can be achieved without resorting to heavy architectures—a crucial property for real-world deployments on mobile or latency-constrained platforms.

3.5.3 Ablation Studies

To measure the impact of specific components within our proposed approaches, we conduct an ablation study by gradually enabling or disabling key features.

Ablations for the HoloVerif Method

We conduct an ablation study to quantify the contribution of each component in the HoloVerif pipeline by incrementally adding: (1) background subtraction, (2) pseudo-labeling, and (3) data augmentations. Table 3.6 presents the results on the MIDV-Holo and MIDV-DynAttack datasets.

Step 1: Exposing Biases via Background Subtraction. Comparing the naive *Direct Classifier* method against its variant trained on background-subtracted frames (*Direct Classifier + BG subtraction*) reveals significant hidden biases: while the F-score on *MIDV-Holo Vanilla* improves by 7 points (90% →97%), the recall on *static-swap* collapses by 75 points (89% →14%). This demonstrates that the *Direct Classifier* exploits template-specific or identity-related cues that vanish once the static background is removed, rather than learning genuine holographic features. The drop in recall on the *Dynamic* attack category (52%→ 25%) also confirms a focus on some dataset biases rather than holographic understanding. However, this preprocessing step is essential for *HoloVerif*, as the model ingests only individual still images, and we stand by the requirement to focus exclusively on the variable holographic signal.

Step 2: Refining Supervision With Pseudo-labels. We next introduce pseudo-labels that mark individual frames from legitimate videos as "non-valid" when insufficient holographic content is detected by the preprocessing stage, creating a more nuanced training signal than binary video-level labels.

Critically, *Pseudo-labels only (no classifier)*—the preprocessing stage of *HoloVerif*—performs poorly: F-score drops to 64% and AUC to 61%, demonstrating that deterministic preprocessing alone cannot replace learned representations. However, when combined with a trained classifier (*HoloVerif without augmentations*), we observe consistent improvements while maintaining a high F-score (96%): *Static* recall increases by 9 points (86% →95%) and *Dynamic* recall by 7 points (25%→ 32%). This confirms that pseudo-labels effectively teach the model to distinguish frames with visible holographic content from those without, rather than simply memorizing video-level associations.

Step 3: Enhancing Generalization With Augmentations. Finally, adding targeted data augmentations designed to simulate diverse capture conditions yields the full *HoloVerif*. The method maintains the same high F-score (96%) while sub-

Table 3.6 – Ablation study of the HoloVerif and Direct Classifier methods. The table compares the impact of background subtraction (BG subtraction), pseudo-labeling, and data augmentations. *Starred methods (`MethodName`*) indicate the variants retained in the main results (Table 3.2).*

	MIDV-Holo (test split)			**MIDV-DynAttack** (full)			**Both**
	Vanilla 120 mixed vids		**Photo Rep.** 20 vids	**Static** 110 vids	**Static-swap** 40 vids	**Dynamic** 90 vids	**Mix** 380 vids
Verification methods ↓	**F-score**	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**	**AUC**
Direct Classifier *	90 ± 4	89 ± 7	69 ± 18	80 ± 15	89 ± 10	52 ± 20	91 ± 3
Direct Classifier + BG subtraction	97 ± 2	98 ± 2	46 ± 11	86 ± 6	14 ± 8	25 ± 11	90 ± 2
Pseudo labels only (no classifier)	64 ± 4	80 ± 14	71 ± 20	87 ± 8	60 ± 22	65 ± 13	61 ± 2
HoloVerif without augmentations	96 ± 2	96 ± 4	43 ± 12	95 ± 4	20 ± 7	32 ± 8	88 ± 3
HoloVerif *	96 ± 2	96 ± 5	59 ± 18	93 ± 5	27 ± 12	61 ± 14	93 ± 2

stantially improving dynamic attack detection: recall increases from 32% to 61%, a gain of 29 points. The overall AUC also improves from 88% to 93%.

The ablation reveals the importance of each component of the *HoloVerif* method: background subtraction removes dataset biases and isolates the holographic signal; pseudo-labels refine supervision to distinguish valid from invalid frames; and augmentations prove essential for robust dynamic attack detection. The improvement brought by augmentations (+29 points on dynamic recall) suggests that further enriching the augmentation strategy—or incorporating synthetic fraud generation—could yield additional gains against dynamic attacks.

Ablations for the WSL Method

Finally, we conducted an ablation study to evaluate the contribution of various components of our WSL approach. Table 3.7 summarizes the results. We compared the following variants: *ImageNet weights (no training)*, in which the backbone was initialized with pretrained ImageNet weights and not fine-tuned (the decision relies on mean pairwise distances between frame embeddings); *WSL without augmentations*, in which data augmentations were completely disabled in the whole pipeline; *WSL only Legit*, in which training was performed exclusively on legitimate samples, without access to attack data; and *WSL* *, which is the reference implementation incorporating augmentations, contrastive loss, and the full training set. Additionally, we evaluate the impact of applying background subtraction as a preprocessing step (methods prefixed by "BG-" in the lower section of Table 3.7).

Necessity of Task-specific Fine-tuning. *Does fine-tuning provide a benefit over generic pretrained features?* Using only the raw mobileVit backbone (*ImageNet weights (no training)*) without fine-tuning on the hologram verification task yields an F-score of only 68%—approaching the minimal F-score at which all samples would be rejected. This confirms that generic pretrained features are insufficient

Table 3.7 – Ablation study of the WSL method. The variants tested include: removal of augmentations (*WSL without augmentation*), training on legitimate samples only (*WSL only Legit*), applying background subtraction (*WSL*) and simply using pretrained weights without fine-tuning (*ImageNet weights*). A variant of the pipeline including a background subtraction preprocessing step is also listed. *Starred methods (`MethodName*`) indicate the variants retained in the main results (Table 3.2).*

	MIDV-Holo (test split)			**MIDV-DynAttack** (full)		
	Vanilla 120 mixed vids		**Photo Rep.** 20 vids	**Static** 110 vids	**Static-swap** 40 vids	**Dynamic** 90 vids
Verification methods ↓	**F-score**	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**
Original pipeline (no background subtraction)						
ImageNet weights (no training)	68 ± 1	90 ± 5	92 ± 10	64 ± 11	55 ± 20	68 ± 13
WSL without augmentations	96 ± 2	96 ± 5	43 ± 26	76 ± 12	16 ± 16	12 ± 12
WSL only Legit	93 ± 5	92 ± 10	90 ± 12	80 ± 10	25 ± 15	4 ± 5
WSL *	95 ± 3	96 ± 5	93 ± 8	80 ± 10	15 ± 9	3 ± 2
Pipeline including a background subtraction preprocess						
BG-ImageNet weights (no training)	67 ± 1	100 ± 0	100 ± 0	98 ± 5	100 ± 0	98 ± 3
BG-WSL	90 ± 4	92 ± 9	60 ± 19	93 ± 5	65 ± 19	41 ± 19

for capturing holographic behavior, and that task-specific fine-tuning is mandatory.

Data Augmentation. *What is the impact of data augmentations?* The triplet loss mechanism, combined with our sampling strategy, relies heavily on augmentations to function effectively. Comparing *WSL without augmentations* to the reference *WSL* shows that augmentations are critical for generalization across different datasets. Specifically, for triplets constructed from Legit (also called "Original") sequences, disabling augmentations effectively nullifies the term $d(a_i, p_i)$ in Equation (3.3), as the anchor a_i and positive p_i become identical. While the model trained without augmentation maintains a high F-score on the MIDV-Holo Vanilla test set—leveraging the explicit video fraud labels provided in the training data—its ability to generalize is reduced. When augmentations are enabled (as they are in *WSL*), the recall for the *Photo Rep.* category improves significantly from 43% to 93%. This improvement is likely attributable to the random crop augmentation, which encourages the network to focus attention on the central region of the image where the face—and the replaced portrait in Photo Replacement attacks—is located. Overall, augmentations prove to be essential for robust generalization.

Training Set Composition. *Are attack samples required to train our approach?* The standard *WSL* configuration uses a balanced mix of non-Legit and Legit samples. However, in real-world scenarios, accessing a diverse set of attack samples is challenging. Therefore, we investigate the viability of training exclusively on Legit samples (*WSL only Legit*). In this experiment, the model checkpoint was selected

based on validation loss calculated solely on legitimate samples, although the final decision threshold was still calibrated using the full validation set (Legit and Non-Legit samples). Training effectively on *MIDV-Holo Vanilla* legitimate samples results in only a minor performance drop (2 points in F-score) compared to the full training setup. Crucially, even this restricted training regime still outperforms the *MIDV-Holo-baseline*, demonstrating the practical value of the method when fraud data is unavailable.

Impact of Background Subtraction. *Does preprocessing with background subtraction improve performance?* Applying background-subtraction preprocessing to our approach (leading to *BG-WSL*) reveals a trade-off: the F-score decreases (95% →90%), but the recall on the *Dynamic* category improves substantially (3%→41%) and *Static-swap* recall increases (15%→ 65%). This suggests that background subtraction further constrains the model to focus on holographic signals. Balancing this improved attack detection with precision remains an open direction for future work.

3.6 Conclusion

This chapter addressed the challenge of verifying holograms on identity documents using only video-level labels. We structured this problem around two distinct use cases established in the introduction: **1) Specific OVD model verification**, which aims to validate the conformity of a target hologram's appearance; and **2) Generic holographic content detection**, a fallback scenario assessing the presence of any holographic behavior when specific data is unavailable.

We proposed and evaluated three data-driven approaches—Direct Classifier, HoloVerif, and WSL—and compared them against the existing MIDV-Holo-Baseline baseline. Our experiments demonstrated that all proposed methods significantly outperform the baseline on the MIDV-Holo benchmark. Critically, the evaluation on the challenging MIDV-DynAttack dataset revealed that methods designed primarily for generic detection (the MIDV-Holo-Baseline baseline and **WSL**) fail against dynamic attacks that mimic holographic behavior. This confirms that verifying *appearance coverage* (i.e., variance) is insufficient; robust systems must also verify *appearance conformity* to distinguish the expected OVD model from generic holographic foils.

We observed complementary failure modes among the successful verification methods. The **Direct Classifier** proved robust against document swapping attacks but was inconsistent against dynamic frauds. Conversely, **HoloVerif** achieved state-of-the-art results on dynamic attacks by leveraging background subtraction and pseudo-labels, yet struggled with alternating template attacks. Notably, **WSL** offers a unique advantage: it is the only method presented that can be effectively trained using legitimate samples alone, making it valuable for scenarios where

fraud data is scarce. Table 3.8 summarizes the strengths and limitations of each approach, highlighting their complementarity.

Several avenues for future work could further enhance these methods. **HoloVerif**, in particular, could benefit from richer training data through synthetic fraud generation. Generative approaches—such as Generative Adversarial Networks (GANs) or diffusion models—could synthesize diverse and realistic attack samples, exposing the classifier to a broader range of forgery patterns. Similarly, advanced augmentation strategies targeting holographic appearance variations could improve robustness without requiring additional real acquisitions.

Despite these advances, all proposed methods share a fundamental limitation: they treat videos as unordered sets of frames rather than coherent temporal sequences. This makes them vulnerable to attacks such as document swapping, where multiple static templates are alternated to simulate variance. Although **WSL** learns a discriminative latent space, its potential remains underutilized by the current decision stage based on mean pairwise distance. Additionally, while qualitative tests suggest some generalization capability, rigorous cross-OVD evaluation remains a necessary step hindered by current data availability.

These limitations point to a clear research direction: learning dynamic OVD representations that capture how their appearance evolves across ordered frames, rather than analyzing each frame independently. The next chapter explores this direction, introducing sequence-level approaches.

Table 3.8 – Summary of frame-wise OVD verification methods. Each approach addresses different aspects of the verification problem, with complementary strengths and limitations.

Method	**Description**	**Strengths**	**Limitations**
MIDV-Holo-Baseline Koliaskina et al. 2023	Heuristic-based detection using pixel-wise hue/saturation variance as proxy for holographic behavior.	No training required Simple and interpretable Effective on basic static printouts	Generic detection only Fails on lighting artifacts Cannot verify specific OVD models Defeated by any holographic content
Direct Classifier (ours)	Naive binary classifier trained with video-level labels applied directly to frames.	Simple baseline Robust to swap attacks Captures diverse signals	Prone to learning dataset biases Fooled by alternate holograms Limited interpretability Uncertain generalization
HoloVerif (Ours)	Background subtraction isolates holographic signal; pseudo-labels and augmentations guide supervised training.	State-of-the-art on dynamic attacks Focuses on holographic content Strong cross-dataset generalization Learns appearance conformity	Vulnerable to swap attacks Requires attack samples for training Background estimation sensitive to alternating templates
WSL (Ours)	Contrastive learning shapes a latent space where dynamic OVD appearances are spread and static imitations cluster.	Trainable with legitimate data only Learns discriminative representations Outperforms baseline on generic detection Generalizes to unseen OVDs	Simplistic decision stage Limited by single-OVD training data Fails on dynamic template attacks Latent space underexploited

CHAPTER 4

SEQUENCE-WISE (TRANSITION-AWARE) OVD REPRESENTATION

Chapter abstract

This chapter presents unpublished exploratory work that extends the frame-level Optically Variable Device (OVD) verification methods introduced in the previous chapter to the sequence domain. The core motivation is to control *dynamic appearance conformity*—which we also call *transition validity*—which cannot be assessed from isolated frames alone. Two complementary approaches are investigated: (1) HoloVerif-Span, a supervised sequence classifier to learn OVD transition characteristics directly from short spans of frames, and (2) a Masked Sequence Modeling (MSM) approach for unsupervised anomaly detection based on reconstruction error. Crucially, both methods are designed to train exclusively on Legit video sequences—i.e., without access to real fraudulent videos. Preliminary results demonstrate the potential of transition-aware verification, particularly in rejecting static-template attacks and simple proxy forgeries. However, challenges remain in detecting sophisticated dynamic attacks that closely mimic genuine hologram behavior. Experiments with synthetic data for the MSM method highlight an interesting potential for detecting subtle temporal anomalies, and suggest that the embedding space learned via masked sequence prediction could be leveraged for assessing the coverage of OVD variety in Legit videos. Overall, these initial findings indicate that modeling temporal dynamics is a promising direction for enhancing OVD verification, although further work is needed to adapt these findings to real-world scenarios.

4.1 Sequence Binary Classifier (HoloVerif-Span)

Method Summary	
short name	HoloVerif-Span
use case(s)	Specific OVD verification.
overview	Supervised sequence-wise binary classifier (span level) + aggregation (video level).
training paradigm	Legit spans (targeted ROI from Legit videos) opposed to non-Legit spans (corrupted Legit spans or spans corresponding to ROIs lacking the target OVD or with a different part).
attack samples	Optional for training and/or calibration.

This section introduces a fifth method called HoloVerif-Span, a variant of the HoloVerif method introduced in Section 3.3.2 that supports sequences as input rather than isolated frames. Leveraging a VideoMAE backbone (Tong et al. 2022), it aims to learn OVD transition characteristics directly from spans (i.e., short sequences of frames), thereby addressing the limitations of the frame-based methods presented in Chapter 3. The task remains the verification of a specific OVD model from a video clip, and the output is a video-level verdict. By extension from the definitions in Section 3.3.2, the terms *Legit / Non-Legit* refer to the final aggregated output at the video level (decision level), while *Valid / Non-Valid* refer to the Region of Interest (ROI) spans processed by the classifier (local signal analysis).

Although our previously presented HoloVerif method can theoretically train on a dataset of Legit videos only—via a pseudo-labeling process that produces Valid and Non-Valid ROIs—experimental results showed a significant performance drop when restricted to such data. Similarly, HoloVerif-Span is designed to be trained exclusively on Legit videos. Non-Valid ROI spans can be generated via contextualized sequence extraction from Legit recordings. Consequently, the model can be trained without access to actual fraudulent videos; real fraud samples can therefore be reserved strictly for testing. Calibration of the decision threshold ideally prompts the use of some fraudulent samples to find the optimal operating point; however, preliminary results indicate that using a default value of 0.5 is feasible.

Figure 4.1 illustrates how spans are extracted from different predefined ROIs (explicitly provided in the MIDV-Holo dataset) and how sequences are preprocessed and eventually corrupted (depending on the ROI source) to provide Non-Valid samples.

VALID SPANS are extracted from ROI containing the target OVD, often positioned over the portrait region (“valid OVD”).

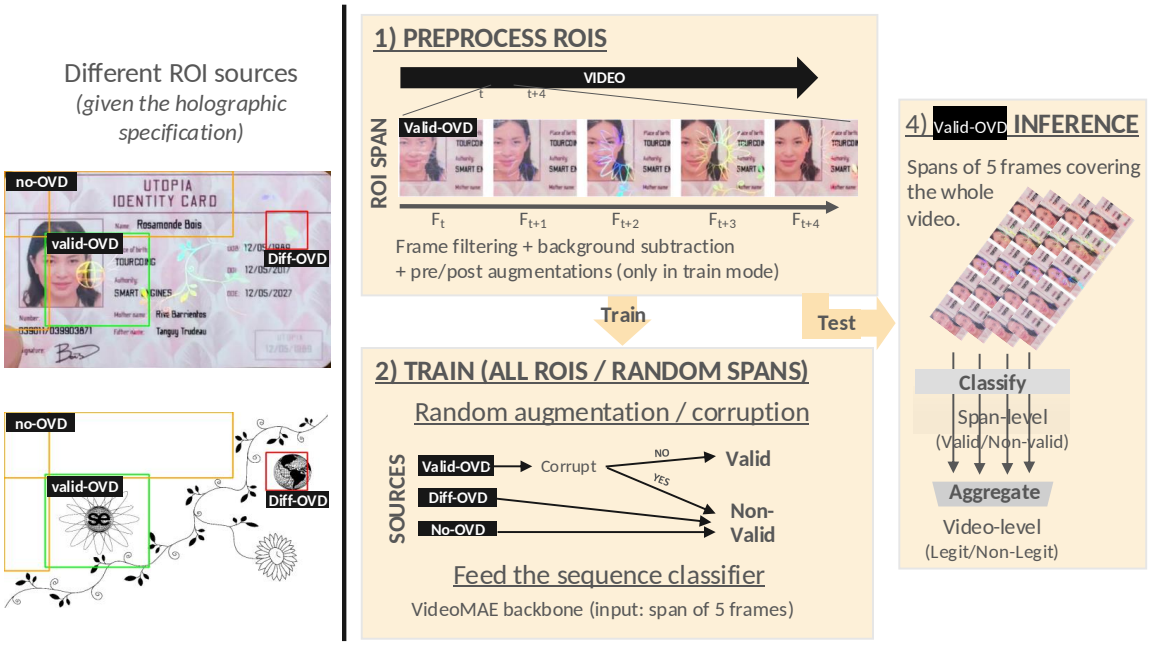

Figure 4.1 – Overview of the full HoloVerif-Span approach (train and test) using visual samples from the MIDV-Holo dataset available for training. The rectified document frame with its corresponding holographic layer mask shows extractable ROIs and their nature (valid/diff/no OVD), which drives the span corruption policy during training.

NON-VALID SPANS are derived from: (i) ROIs without a hologram ("no OVD"), (ii) ROIs from the same or different documents carrying a *different* OVD than the target (model mismatch, "diff OVD"), and (iii) crafted manipulations of Legit videos, such as shuffles, hard cuts, copy/loops, and swap-style mixes (Non-Valid samples generated from "valid OVD").

The following sections detail the steps of the proposed pipeline.

4.1.1 Preprocessing

Spans (sequences of 5 frames cropped to a specific ROI) are extracted from full rectified videos sampled at 15 fps. To improve data quality, the following preprocessing pipeline—slightly differing between train and test modes—is applied in order:

1. **Tracking consistency filtering:** Structural Similarity (SSIM) (Wang et al. 2004) and pixel-wise frame differences are used to detect tracking failures or misalignments (e.g., ROI jumps, dewarping glitches). Spans that violate consistency thresholds are discarded.

2. **Pre-augmentations (TRAIN MODE ONLY):** These are applied identically to all samples, regardless of the ROI source, preventing the classifier from relying on trivial global brightness or saturation cues. Techniques include:

   - Color jitter (brightness, contrast, saturation, hue) to emulate changing illumination;
   - Mild Gaussian blur to approximate motion blur and defocus;
   - Random equalization, autocontrast, sharpness adjustments, posterization, and small erased patches to mimic sensor/compression artifacts or partial occlusions;
   - Synthetic shadows and glare via Kornia (Riba et al. 2020), using `RandomPlasmaShadow` and `RandomPlasmaContrast` operators from Nicolaou et al. 2022;
   - A custom `RandomPlasmaQuickdrawBrightness` transform that modulates brightness using fractal-like fields masked by silhouettes from Quick!Draw! (Jongejan et al. 2017), producing localized bright/dark blobs reminiscent of complex lighting artifacts.

3. **Background subtraction:** A local background is estimated over a short temporal window centered at each frame (rather than over the full video) and subtracted to attenuate static template and face content. This approach is robust to slow drift, minor pose changes, and local misalignments. A hue/saturation enhancement factor is also applied to boost the signal of potential OVD signal.

4. **Post-augmentations (TRAIN MODE ONLY):**

   - For Valid spans—representing genuine behavior—we apply a conservative resize to 224×224 pixels without strong geometric distortions to preserve thin OVD patterns.
   - For Non-Valid spans, we apply random resized crops to 224×224 pixels, flips, small rotations, additional RandomPlasma-based effects, and photometric jitter.

4.1.2 Training

Label Balancing via Corruption. This key step enables training in the absence of Non-Legit videos. It augments data diversity by generating Non-Valid spans from Valid ones using two corruption strategies:

- *Frame repetition*: Select 1–3 unique frames and repeat them (with subtle per-frame jitter) to fill the 5-frame sequence, simulating buffer stalls, simple loops, or nearly static content.
- *Mixed corruption*: Randomly choose between (i) partial repeats (random positions reuse previous frames), (ii) shuffles with repeats (70% unique, 30% repeated frames, shuffled), and (iii) segment reordering (2–4 contiguous subsequences are permuted).

These sequences retain a hologram-like appearance in individual frames, but their temporal evolution is deliberately non-physical.

Classifier Architecture. For temporal modeling, we rely on a Transformer-based video encoder derived from VideoMAE (Tong et al. 2022). VideoMAE is a masked autoencoder for video; during pre-training, spatio-temporal patches of the input clip are randomly masked, and the network is trained in a self-supervised manner to reconstruct missing content from visible patches. This forces the encoder to learn compact representations capturing both appearance and motion without manual labels. We utilize the publicly available "small" variant ($\approx$ 22M parameters), pre-trained with this objective and fine-tuned for action recognition on Something-Something V2 (Goyal et al. 2017). We adapt this model to accept spans of 5 frames and output two classes (Valid / Non-Valid).

Training Parameters. We experimentally determined the following defaults:

- **Loss:** Cross-entropy between predicted logits and target labels.
- **Optimizer:** AdamW with a learning rate of 10^{-4}.
- **Scheduler:** Cosine decay using a warm-up phase covering 10% of the estimated total training steps.
- **Batch size:** 64 spans.
- **Epochs:** 3 per fold, which proved sufficient to reach a stable validation loss.

4.1.3 Inference and Aggregation for Video-level Decision

The VideoMAE classifier outputs a confidence score per ROI span. Similar to the HoloVerif method, an aggregation step is required to obtain the final verdict at the video level.

First, spans corresponding to the checked ROI (typically over the portrait) are extracted from the rectified video using a non-overlapping window of 5 frames. The preprocessing steps (excluding training-specific augmentations) are applied

Table 4.1 – Preliminary results for HoloVerif-Span. Metrics aggregate the test results over the 5 folds.

	MIDV-Holo				MIDV-DynAttack		
	Originals	**Static**	**Photo Rep.**	**Other ovd**	**Static**	**Static-swap**	**Dynamic**
	Specificity	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**	**Recall**
HoloVerif Span (ours)	0.99	0.47	0.54	0.96	0.81	0.54	0.79

before feeding the spans into the classifier to obtain a fraud score $s_t \in [0, 1]$. These scores are then aggregated by averaging to produce $\bar{s}$. Since the classifier is designed as a non-validity estimator, high scores indicate Non-Valid content. Given an operating threshold τ (supposed to be calibrated on a validation split by maximizing the F-score), the final prediction is:

$$\text{decision} = \begin{cases} \text{Non-Legit}, & \text{if } \bar{s} \geq \tau, \\ \text{Legit}, & \text{otherwise.} \end{cases}$$

For the MIDV-Holo and MIDV-DynAttack datasets, all attack categories (e.g., copy without holo, pseudo holo copy, photo replacement, static, and dynamic families) are treated as Fraud and should ideally produce $\bar{s} \geq \tau$. Only MIDV-Holo "Legitss" are considered Legit and should remain below the threshold.

4.1.4 Preliminary Results and Analysis

This section presents preliminary results obtained with the HoloVerif-Span method. These results are not directly comparable to those in Chapter 3 as the protocol is not yet fully aligned; specifically, no calibration using a validation set was performed, and the optimal F-Score metric is not provided. Additionally, training and evaluation were conducted exclusively on the Identity Document portion of the MIDV-Holo and MIDV-DynAttack datasets. Nevertheless, the concepts presented in this fourth chapter—most notably, training without Non-Legit samples—are promising exploratory avenues.

Results in Table 4.1 follow the identity-disjoint K-fold scheme of Section 2.2.4, applied to the MIDV-Holo and MIDV-DynAttack datasets. Additionally, Figure 4.2 summarizes the distribution of per-video fraud scores across all folds and categories. For each category, the figure reports how many videos lie above the default 0.5 threshold (thus classified as Non-Legit). Unlike approaches in previous chapters, no calibration was performed over a validation set, and no fraudulent videos were seen during training or validation.

Legits: Very Low False-alarm Rate. For MIDV-Holo Legits, almost all videos yield fraud scores well below the threshold. Only a tiny fraction of Legits are

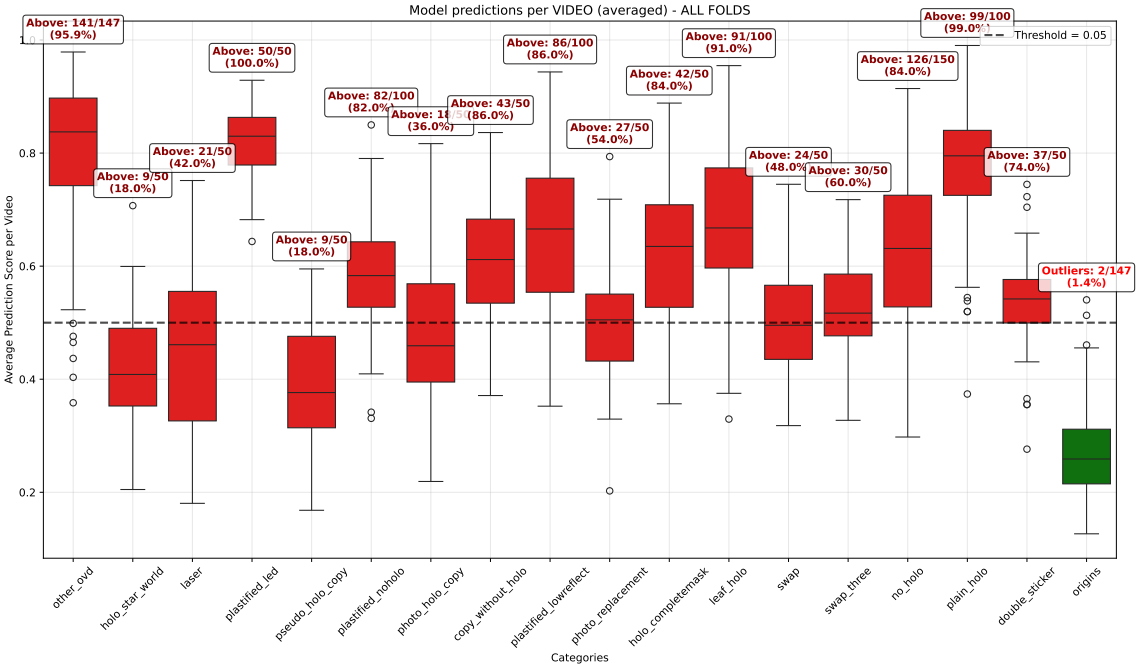

Figure 4.2 – Per-video fraud scores (averaged over all 5-frame windows) across all folds and categories. Each bar corresponds to a category from MIDV-Holo or MIDV-DynAttack (Legits, background ROIs, static-template attacks, swap attacks, and dynamic-template attacks). The horizontal line marks the operating threshold τ (set to 0.5 in the absence of calibration on a validation set with Non-Legit samples); videos with scores above this line are classified as Fraud. Annotations of the form "Above: a/b (p%)" indicate how many videos in each category are flagged as Fraud. Legits cluster well below the threshold (few outliers), while most static/no-OVD and many simple proxy attacks are correctly rejected. Advanced dynamic and swap-based attacks remain only partially detected, illustrating both the benefits and limitations of transition-based verification.

misclassified as Fraud. These outliers typically correspond to unusual capture conditions (e.g., extreme glare, strong saturation in the ROI) where the enhanced signal resembles some of the synthetic temporal negatives. Overall, this confirms that the sequence classifier preserves a high acceptance rate for genuine holograms.

Static No-ovd and Simple Proxy Attacks: Largely Rejected. Static-template attacks that lack holograms or rely on basic lighting proxies are usually assigned high fraud scores. This includes: MIDV-Holo **copy without holo** and paper copies; MIDV-DynAttack "NL-Paper" (unlaminated prints) and "NL-Plastic" (laminated copies); and various LED/laser-based attacks where dynamics are confined to small, highly localized spots differing from global holographic behavior. For these families, a large majority of videos are correctly classified as Fraud. This is consistent with the training design, as `other` ROIs and temporally corrupted OVD sequences cover many such "no real hologram" and "lighting-only" patterns.

Static Swaps: Partially Detected. Swap attacks using two or three printed static templates (`swap`, `swap_three`) prove more challenging. When template changes are infrequent within a 5-frame window, some windows may mimic a plausible short hologram trajectory, leading to modest fraud scores. While faster swapping increases detection rates, performance remains far from the ideal 100%. This is expected: the model only sees short clips and is not explicitly trained to recognize discrete template jumps beyond what is implicitly covered by synthetic temporal corruption.

Handcrafted Dynamic Holograms: Mixed Performance. For MIDV-DynAttack dynamic-template attacks ("plain_holo", "leaf_holo", "double_sticker", "complete_mask", and "star_and_world"), the classifier exhibits mixed behavior. A non-negligible fraction of videos per category are correctly flagged, especially when the added holographic material produces dynamics that differ markedly from the reference OVD (e.g., incorrect scale, location, or obvious shape mismatch). However, a significant proportion remains below the threshold, particularly for attacks where the dynamic overlay roughly matches the expected location and gross behavior of the original hologram. In these borderline cases, the model appears to focus more on the presence of holographic-like temporal variability than on the fine-grained conformity of individual appearances.

Photo Replacement: Still Problematic. As observed with frame-wise methods, photo replacement in MIDV-Holo remains one of the hardest scenarios. Although the portrait is fraudulently replaced, part of the genuine hologram often remains visible in the ROI and exhibits correct dynamics. Consequently, the sequence classifier frequently assigns low fraud scores, incorrectly accepting such attacks as Legit. Because partially observable OVD signals are common in legit videos,

the model has limited incentive to treat these cases as anomalous. This underlines the intrinsic difficulty of photo replacement attacks, which may require adding a *spatial coverage* component to the verification pipeline, specific heuristics for face picture ROIs, or tighter ROI focus.

4.1.5 Discussion and Limitations

Designing a transition-aware verifier under the constraint of using only Legit videos is intrinsically difficult. The model must learn what *invalid* temporal behavior looks like without ever seeing real fraudulent trajectories. In practice, this forces us to manufacture negative sequences by (i) sampling ROIs where the OVD is absent and (ii) corrupting genuine OVD clips through frame repetition, shuffling, and segment reordering. While these synthetic sequences are sufficient to teach the classifier to reject obviously non-holographic or grossly inconsistent dynamics, they only partially cover the space of realistic forgeries.

Empirically, the resulting model behaves as expected at the two extremes: it performs well when the attack is clearly outside the learned manifold of the target OVD, e.g. no-OVD regions, pure lighting proxies, or dynamic overlays whose shape, position, or global dynamics differ markedly from the reference hologram. In these regimes, the synthetic negatives provide a useful training signal and the learned decision boundary is relatively robust.

The most challenging regime is that of *similar* OVDs and near-miss forgeries. Handcrafted dynamic attacks in MIDV-DynAttack that approximate the correct location and overall color behavior, or swap attacks that alternate between several plausible hologram states, still lie close to the Legit manifold. Our current synthetic sequences, based on simple repetition and permutation of frames, are too coarse to populate this region of the space: they generate clearly abnormal temporal patterns, but do not emulate the subtle, structured deviations present in sophisticated forgeries. As a result, the classifier has limited incentive to push the decision boundary tighter around the genuine trajectories.

Addressing this limitation will likely require *harder* negative examples. Two complementary directions appear promising: (i) enriching the dataset with more realistic, carefully engineered frauds that approximate the true OVD while remaining invalid (for example, by modifying only part of the hologram layer), and (ii) leveraging generative models (Generative Adversarial Networks (GANs) , diffusion, or sequence-level generators) trained to produce adversarial hologram evolutions that are close to but inconsistent with genuine behavior. Such models could be used either to augment the training set with near-boundary negatives, or within an adversarial training loop where the generator explicitly seeks sequences that the verifier misclassifies. In both cases, the goal is to populate the "gray zone" between obviously wrong dynamics and genuine OVD behavior, so that Transition Validity can be learned and enforced at a much finer granularity.

4.2 Sequence Anomaly Detection via Masked Sequence Modeling (MSM)

Method Summary	
short name	Masked Sequence Modeling (MSM)
use case(s)	Specific OVD verification.
overview	Masked Sequence modeling + reconstruction error.
training paradigm	Generative through masked input reconstruction.
attack samples	Optional for training and calibration.

This section presents a sixth method, referred to as Masked Sequence Modeling (MSM), which utilizes masked sequence prediction for unsupervised anomaly detection based on reconstruction error. Designed to train exclusively on Legit video sequences without requiring access to fraudulent samples—akin to the HoloVerif-Span—this approach models the temporal evolution of OVD appearances. By learning the dynamics of genuine OVDs, it aims to identify deviations from these learned patterns as indicators of potential forgeries. We first review relevant prior work on sequence learning and anomaly detection, then detail the proposed architecture. Finally, we present preliminary results and discuss limitations and future directions.

4.2.1 Related Work

Sequence-to-sequence Architectures. Several architectural families have been proposed for sequence modeling in machine learning. Their primary characteristic of interest in this context is the ability to produce a contextualized representation for each element by integrating information from its neighbors. Such representations facilitate various downstream tasks, including classification, regression, and—in our case—generation. We distinguish three main families: autoregressive, bidirectional, and global models.

Autoregressive models, such as Recurrent Neural Networks (RNNs), Long Short-Term Memorys (LSTMs) (Hochreiter and Schmidhuber 1997), Gated Recurrent Units (GRUs) (Chung et al. 2014), and decoder-only Transformers (Vaswani et al. 2017), process sequences unidirectionally, predicting each element based solely on past context. These are well-suited for tasks like language modeling or time-series forecasting where causality is paramount. Transformer-based models have achieved state-of-the-art results across many domains due to their ability to capture long-range dependencies via self-attention. However, they are compu-

tationally expensive (scaling quadratically with sequence length) and typically require large datasets for effective training. RNN architectures offer a lightweight alternative. LSTMs mitigate the vanishing gradient problem inherent in vanilla RNNss through gating mechanisms, maintaining dependencies over longer horizons. GRUs simplify the LSTM cell by merging the forget and input gates, resulting in fewer parameters and faster training while maintaining comparable performance.

Bidirectional models, such as bidirectional LSTMs (Graves and Schmidhuber 2005) or Transformer encoders (e.g., BERT Devlin et al. 2019), attend to both past and future contexts when encoding each element. This is advantageous for tasks like sequence classification or masked prediction where the full context is available simultaneously. We also note that Convolutional Neural Network (CNN)-based architectures have been proposed for sequence modeling (Bai et al. 2018), using temporal convolutions to capture local dependencies. Such models may present interesting properties in terms of efficiency and contextual representation, and could be explored in future work.

Finally, *global models* process and encode the entire sequence holistically, often without explicit temporal ordering or element-wise contextualization. Autoencoders (Hinton and Salakhutdinov 2006) exemplify this category, learning to compress and reconstruct sequences. Variational Autoencoders (VAEs) (Kingma and Welling 2013) introduce a probabilistic framework, enabling the generation of new sequences by sampling from a learned latent space. However, their application to pure sequence data is less standard than other architectures, and variable-length sequences can complicate training. Diffusion models have recently been proposed for sequence generation (Hoogeboom et al. 2022). Although exhibiting promising capabilities for anomaly detection in time series (Chen et al. 2023) and medical imaging (Iqbal et al. 2024), their high computational cost and complexity currently render them less suitable for our constraints. We leave their exploration to future work.

For our application—involving short sequences (tens of frames), modest data volumes, and a requirement for efficient inference—we adopt a bidirectional LSTM-based architecture. This choice balances modeling capacity with computational costs, and our experiments confirm that LSTMs suffice to capture the temporal regularities present in our data.

Masked Sequence Modeling. Masked prediction has emerged as a dominant paradigm for self-supervised representation learning. The seminal BERT model (Devlin et al. 2019) demonstrated that randomly masking tokens in a sentence and training a Transformer encoder to reconstruct them from bidirectional context yields representations that transfer robustly to downstream tasks. This approach has been extended to continuous domains: masked autoencoders for images (He et al. 2022) mask random patches and reconstruct pixel values, while VideoMAE (Tong et al. 2022) extends this to spatio-temporal video patches. Unlike token-

based masked language models that predict discrete vocabulary indices, our approach operates in a continuous embedding space. This requires the model to regress dense embedding vectors rather than selecting from a fixed set of tokens.

Sequence Anomaly Detection. The key insight underlying masked prediction is that accurate reconstruction necessitates an understanding of the data's underlying structure. A model capable of predicting masked content from context must have learned meaningful patterns and regularities. Conversely, when presented with out-of-distribution samples that violate this structure, the model's predictions degrade, resulting in elevated reconstruction error. This property makes masked prediction a natural foundation for anomaly detection: by training on normal data, samples with high reconstruction error can be flagged as anomalous. The use of reconstruction error for anomaly detection has a substantial history in machine learning. Models trained to compress and reconstruct normal data inevitably produce higher loss on anomalous inputs. This principle has been successfully applied to images (An and Cho 2015), time series (Malhotra 2016), and video (Hasan et al. 2016).

Regarding techniques suitable for OVD verification, two main paradigms can be distinguished based on the reconstruction domain:

- **Image-space reconstruction**: The model reconstructs raw pixels or frames directly. While intuitive, this approach is ill-suited to our application: the loss function would force the model to dedicate capacity to reconstructing the document background, the bearer's face, and sensor noise—none of which carry information about OVD authenticity. Anomaly detection would consequently be dominated by irrelevant errors in these regions.

- **Embedding-space reconstruction**: The model operates on learned representations rather than raw pixels. By projecting frames into a compact latent space before sequence modeling, the approach focuses on features relevant to the learned embedding (ideally holographic dynamics) while filtering out low-level noise. The reconstruction loss then measures deviations in this structured space.

The embedding-space paradigm is particularly appealing for OVD verification. It focuses the model on meaningful feature trajectories and removes the burden of high-fidelity image generation, simplifying the learning task. Furthermore, it allows leveraging pretrained frame encoders. In this paradigm, reconstruction-based methods can naturally employ masking at verification time. For bidirectional models, masking forces the model to leverage context from both past and future frames to reconstruct missing elements, preventing trivial identity mappings (Fu and Xue 2022). In all cases, the reconstruction error serves as the anomaly score, which requires calibration on a validation set to determine an optimal operating threshold.

4.2.2 Proposed Method

Our approach operates in two stages: (1) learning (or adapting) a frame encoder—hereafter referred to as the *frame projector*—that maps video frames to a structured embedding space, and (2) training a sequence model to predict masked portions of these embedding trajectories. At inference, anomalous dynamics are detected through elevated reconstruction error. Figure 4.3 illustrates this proposed pipeline. Please note that this approach does not rely on background subtraction or enhancement, as the frame projector is expected to learn to focus on holographic features.

We considered several strategies for the frame projector: training it from scratch (or fine-tuning) using either exclusively Legit data or a mix of Legit and Non-Legit samples (potentially including diverse Legit OVDs), or reusing an off-the-shelf pre-trained model (e.g., DinoV2 (Oquab et al. 2023)). The latter option presents the specific challenge of directing the network's attention toward valid OVD content rather than portrait or background features, which are typically more salient in generic pre-trained models.

In our preliminary experiments, we focused on training the projector from scratch using exclusively Legit data containing a single synthetic OVD model. To essentially capture holographic appearance, we adapted the Weakly Supervised Learning (WSL) approach described in Section 3.4.

Training the Projector (stage 1)

Starting From the WSL Approach. We construct training triplets similarly to our WSL method presented in Section 3.4.2, utilizing temporal proximity as the source of supervision. Specifically, triplets are formed as follows:

- **Anchor and Positive:** Temporally adjacent frames $(t, t+1)$ with independent data augmentations;
- **Negative:** A temporally distant frame $(t + 5)$, encouraging the embedding to capture temporal variation.

This sampling strategy teaches the encoder to map augmented versions of the same instantaneous holographic state to similar representations, while separating embeddings of distinct states observed at different times.

Preventing Representational Collapse via Smoothness Regularization. For effective out-of-distribution detection, the encoder must map non-conforming samples to regions of the embedding space distinct from legitimate samples. However, we identified a critical failure mode when training the WSL approach exclusively on legitimate sequences as described in Section 3.4. Because negative samples are drawn from a relatively short temporal window (typically 5 frames

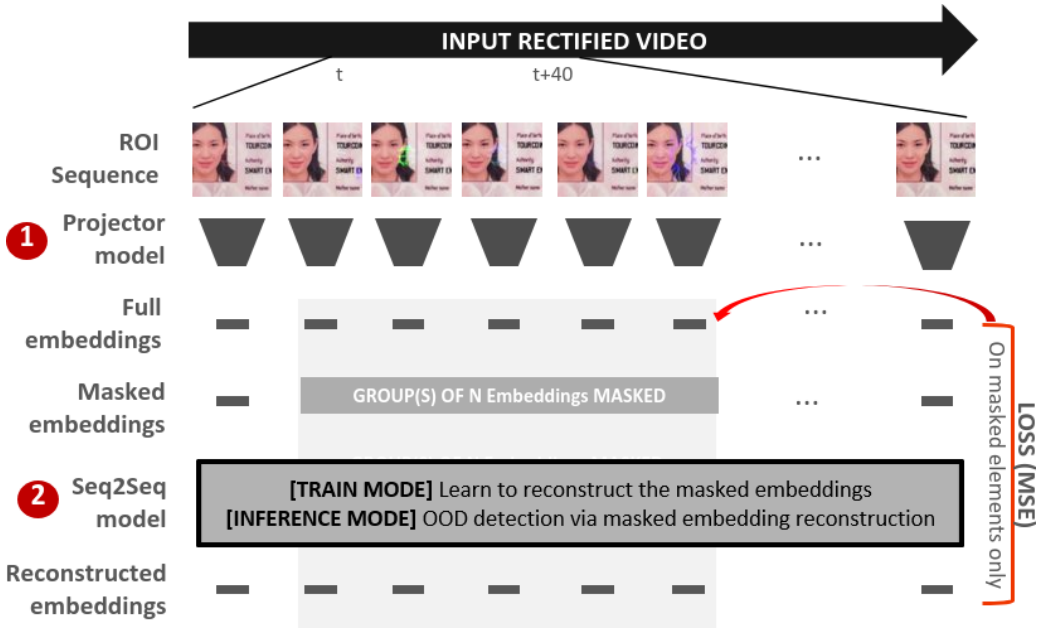

Figure 4.3 – Our generic MSM approach illustrated. The *Projector* refers to any model capable of extracting holographic features (e.g., pretrained DinoV2, custom WSL). The Seq2seq model refers to any model capable of catching transition behavior and reconstructing sequences from masked elements (e.g., LSTM, Transformer encoder).

apart), visually distinct phases of the OVD that never occur in close temporal proximity are never explicitly contrasted as mutual negatives. Consequently, the encoder may fail to push these distinct phases apart, causing them to collapse into overlapping regions of the latent space.

We visualized this phenomenon using a synthetic dataset designed to mimic smooth OVD transitions. The generated samples consist of portrait regions overlaid with blends of 4 distinct Quick, Draw! shapes serving as visual references. Ideally, the projector should organize these inputs into 4 well-separated clusters corresponding to the distinct references. However, as illustrated in Figure 4.4 (left), the naive WSL approach merges visually distinct samples into only 3 clusters, confirming the collapse.

To address this, we integrated a *smoothness loss* into the training objective. This regularization term minimizes the cumulative distance along sequence trajectories in the embedding space. Generally, the triplet loss encourages separation between temporally distant samples, while the smoothness loss encourages gradual transitions between adjacent ones. In practice, these partially opposing constraints produce well-structured embedding spaces where distinct appearances form separate clusters connected by continuous paths (Figure 4.4, right). Effectively, if the path from state A to B to C is constrained to be smooth and continuous, A and C are implicitly prevented from occupying the same location.

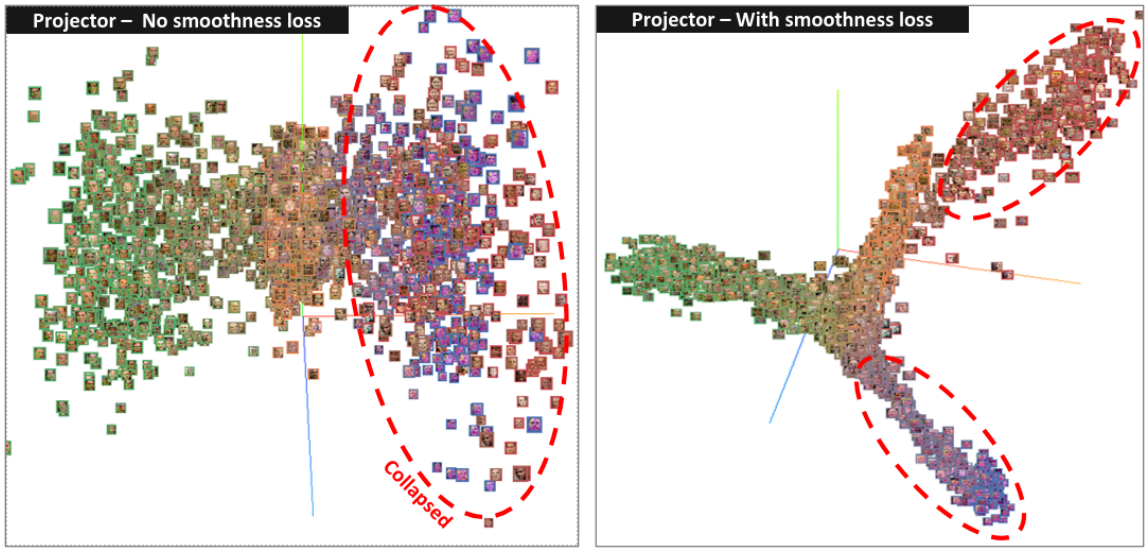

Figure 4.4 – Visualization of the embedding space learned by the projector model using synthesized samples generated from 4 distinct visual references (alpha-blended). On the left, training without the smoothness regularization term leads to a representational collapse where distinct phases are merged into only 3 clusters. On the right, adding the smoothness loss resolves this issue, producing 4 distinct clusters connected by continuous transitions between temporally adjacent samples.

We acknowledge that this solution is pragmatic rather than fundamental. While the smoothness loss effectively mitigates collapse in our experiments, it does not strictly guarantee that all distinct holographic patterns are mapped to disjoint regions. A more robust approach might involve explicit modeling of OVD identity—perhaps by leveraging foundation models (e.g., DINOv2) focused on holographic content via background subtraction, by incorporating physical constraints like angular velocity, or simply by incorporating diverse OVD models in the training. Nevertheless, we leave these extensions for future work, focusing here on assessing the efficacy of the proposed smoothness constraint on the learned projection space.

Final Configuration. The projector model is trained using a triplet margin loss with temporal sampling. From each 40-frame video sequence, we sample 15 consecutive frames. For each 15-frame sample, three triplets are constructed:

- **Anchors**: frames at positions $t \in \{0, 5, 9\}$;
- **Positives**: adjacent frames at positions $t + 1$ (same appearance, different augmentation);

- **Negatives**: distant frames at positions $t + 5$ (different appearance).

The total loss for each sample combines the triplet losses with smoothness regularization:

$$\mathcal{L} = \frac{1}{3}\sum_{k=1}^{3}\mathcal{L}_{\text{triplet}}^{(k)} + \lambda \cdot \mathcal{L}_{\text{smooth}} \tag{4.1}$$

where each triplet loss uses mean squared error as the distance function with margin $m = 0.5$:

$$\mathcal{L}_{\text{triplet}} = \max\left(0,\ d(\mathbf{a},\mathbf{p}) - d(\mathbf{a},\mathbf{n}) + m\right), \quad d(\mathbf{x},\mathbf{y}) = \frac{1}{D}\sum_{i=1}^{D}(x_i - y_i)^2 \tag{4.2}$$

and the smoothness loss penalizes large jumps between all 14 consecutive embedding pairs:

$$\mathcal{L}_{\text{smooth}} = \frac{1}{14}\sum_{t=1}^{14} d(\mathbf{e}_t, \mathbf{e}_{t-1}) \tag{4.3}$$

In our current state of development, the projector model produces 3-dimensional embeddings—convenient for visualization—but higher dimensions may be preferred in practice for increased capacity. Training uses 3,680 video sequences of 40 frames each, with smoothness weight $\lambda = 0.5$.

Stage 2: Masked Sequence Prediction

We freeze the frame encoder and train a sequence model to reconstruct masked embedding sequences. Given a sequence of $T = 40$ frame embeddings $\{\mathbf{e}_t\}_{t=1}^{T}$ with $\mathbf{e}_t \in \mathbb{R}^D$, we randomly select $n = 2$ non-overlapping spans of $L = 5$ consecutive frames and set their values to zero, producing the masked sequence $\{\tilde{\mathbf{e}}_t\}$. The model objective is to reconstruct embeddings $\{\hat{\mathbf{e}}_t\}$ from this masked input that are close to the original embeddings $\{\mathbf{e}_t\}$ at the masked positions. Exploring different strategies for mask selection (random spans, fixed windows, varying lengths) and masking values (zeros, special values, noise) is left for future work.

Architecture. The model is a bidirectional LSTM with input dimension $D = 3$, hidden dimension $H = 64$, $N = 4$ layers, and dropout $p = 0.1$. A linear output layer maps the concatenated forward-backward hidden states ($2H = 128$ dimensions) back to D dimensions.

Loss Function. We minimize the MSE loss computed only on masked positions:

$$\mathcal{L}_{\text{seq}} = \frac{1}{n \cdot L}\sum_{t \in \mathcal{M}} \|\hat{\mathbf{e}}_t - \mathbf{e}_t\|_2^2 \tag{4.4}$$

where $\mathcal{M} \subset \{1, \ldots, T\}$ denotes the set of masked indices with $|\mathcal{M}| = n \cdot L = 10$.

Training. We use AdamW with learning rate $\eta = 10^{-4}$, batch size $B = 64$, and gradient clipping at norm 1. Random temporal flipping (sequence reversal with $p = 0.5$) serves as data augmentation. Training runs for 1600 epochs on the encoder's validation set to prevent overfitting to encoder-specific artifacts.

Inference Pipeline

At inference, the model operates as an anomaly detector by systematically masking and reconstructing segments of the input sequence. Given a sequence of $T = 40$ embeddings, we apply a non-overlapping sliding window strategy with a mask size of $W = 5$ and a stride of $S = 5$. To mitigate boundary artifacts, we exclude the first and last $P = 5$ frames from the masking process; these frames serve solely as temporal context for the bidirectional encoder. This procedure results in $K = 6$ masked windows coverage positions 5 through 34. For each window, the corresponding embeddings are set to zero, and the model predicts the missing values based on the visible context. Because the windows do not overlap, each embedding within the evaluation range is reconstructed exactly once.

The final anomaly score $\mathcal{S}$ is computed as the Mean Square Error (MSE) between the original input embeddings $\mathbf{e}_t$ and their reconstructions $\hat{\mathbf{e}}_t$:

$$\mathcal{S} = \frac{1}{T - 2P} \sum_{t=P}^{T-P-1} \|\hat{\mathbf{e}}_t - \mathbf{e}_t\|_2^2 \tag{4.5}$$

A high score indicates that the observed trajectory deviates from the legitimate dynamics learned during training, signaling a potential forgery.

4.2.3 Experimental Validation on Synthetic Data

We propose a preliminary evaluation of the MSM on a synthetic dataset designed to mimic holographic state transitions. This controlled setting allows us to validate the method's ability to detect anomalies in temporal dynamics without the confounding factors present in real-world data. We detail the dataset generation process, fraud variants, and present quantitative and qualitative results.

Synthetic Dataset Generation. To train the encoder without requiring real holographic documents, we generate synthetic sequences that simulate holographic state transitions. The generation process works as follows:

1. **Holographic states.** We sample 3 shapes from the Quick, Draw! dataset (Jongejan et al. 2017) (denoted A, B, C) and apply random augmentations

(rotation, dilation, color jitter) to simulate diverse holographic patterns, this augmented quickdraw will not change throught the process.

2. **Transition constraints.** We define a linear transition graph: A $\leftrightarrow$ B $\leftrightarrow$ C, meaning direct transitions between A and C are *not* allowed. This simulates the physical constraint that holographic states change gradually under continuous camera motion.
3. **Sequence generation.** For each 40-frame sequence:
   - A background face image is sampled from the LFW dataset Learned-Miller et al. 2016;
   - Starting from a random state (e.g., A), we transition to an adjacent state (e.g., B) over 3–8 frames;
   - Each transition frame is generated by alpha-blending the two shapes according to an interpolation curve (linear, quadratic, or cosine easing);
   - The blended shape is overlaid onto the background;
   - The process continues, alternating between adjacent states, until 40 frames are generated.
4. **Frame labels.** Each frame stores the blending ratio of each shape (e.g., $\{\text{A} : 0.3, \text{B} : 0.7\}$), this is used only for visualization—not during training.

Fraud Variants. To evaluate anomaly detection on the test set, we generate three categories of fraudulent sequences alongside legitimate ones:

- **Wrong Quick, Draw!**: sequences using different shapes (D, E, F, ...) not matching the expected template, testing appearance anomaly detection;
- **Wrong transitions**: sequences violating the A $\leftrightarrow$ B $\leftrightarrow$ C constraint (e.g., direct A $\rightarrow$ C), testing invalid path detection;
- **Sharp transitions**: sequences with abrupt phase changes lacking smooth interpolation, testing temporal discontinuity detection.

Note that some randomly generated sequences may accidentally follow valid paths. These fraud variants are used only for evaluation—training is performed exclusively on legitimate sequences.

Dataset Split on Identity. The dataset construction proceeds in two stages. First, 5,750 unique identities are partitioned into training (64%), validation (16%), and test (20%) splits, ensuring no identity appears in multiple folds. Second, sequences are generated for each identity: one legitimate 40-frame sequence per identity across all splits, plus—for test identities only—three fraudulent sequences

Table 4.2 – Synthetic dataset split by identity. Each identity yields one legitimate sequence. For test identities, three additional fraudulent sequences are generated-one per attack type (wrong appearance, random transitions, sharp transitions).

		Number of Sequences		
Split	**Identities**	**Legitimate**	**Fraudulent**	**Total**
Training (64%)	3,680	3,680	—	3,680
Validation (16%)	920	920	—	920
Test (20%)	1,150	1,150	3,450	4,600
Total	5,750	5,750	3,450	9,200

Table 4.3 – Detection performance on synthetic data. The decision threshold is calibrated to fix the False Positive Rate (FPR) at 1% on legitimate samples. We report the Recall (detection rate) for each specific attack type.

	Legit (Reference)	**Wrong Appearance**	**Random Transitions**	**Sharp Transitions**
	Specificity (1-FPR)	**Recall**	**Recall**	**Recall**
Masked Embedding	99.0%	99.8%	85.0%	100.0%

corresponding to each attack type (wrong appearance, random transitions, and sharp transitions). This results in 9,200 total sequences, as detailed in Table 4.2.

The training set (3,680 identities) is used exclusively to train the encoder on legitimate sequences. The validation set (920 identities) serves to train the sequence-to-sequence model on legitimate sequences and for model selection. The test set (1,150 identities) is reserved for final evaluation, comprising 1,150 legitimate sequences and 3,450 fraudulent sequences ($1{,}150 \times 3$ attack types).

This identity-based splitting strategy follows the same protocol as described in Section 3.5.1, ensuring strict separation between splits. The encoder is trained and selected using only the training and validation sets respectively, while the sequence-to-sequence model is trained exclusively on the validation set with legitimate sequences. Final performance metrics, including reconstruction error on masked samples, are reported on the held-out test set.

Quantitative Results. We evaluate the performance of our method by treating the reconstruction error as an anomaly score. As this is an exploratory study on synthetic data we calibrate the decision threshold *a posteriori* on the test set to target a fixed False Positive Rate (FPR) of 1%. This corresponds to setting the threshold at the 99th percentile of the reconstruction error distribution for legitimate sequences. Ideally, in a full deployment scenario, this calibration would be performed on a separate validation split. Using this operating point, we report

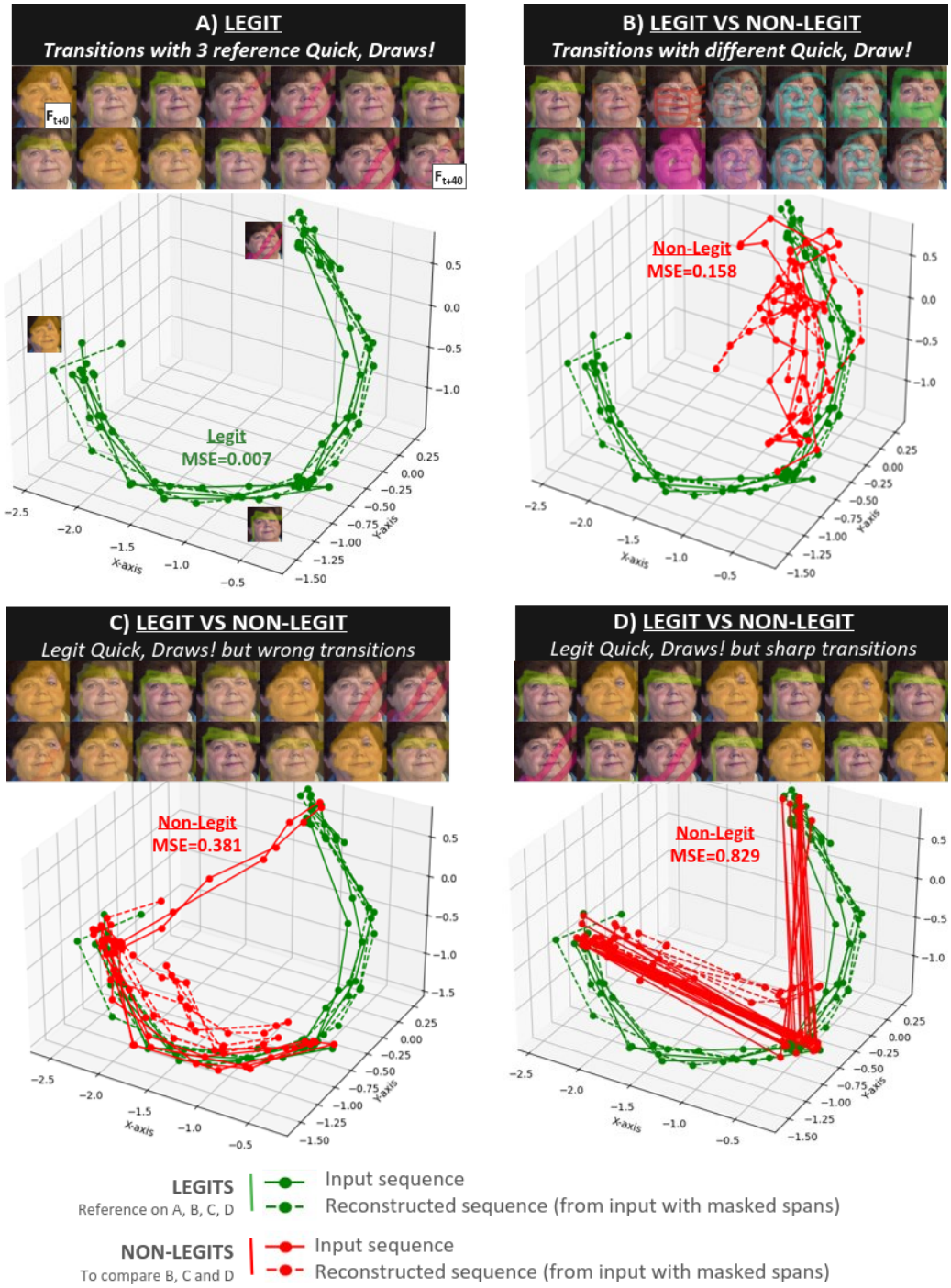

Figure 4.5 – Visualization of embedding trajectories in the latent space for anomaly detection. Solid lines represent the input sequences, while dashed lines show the model's reconstruction of masked segments (used to compute the anomaly score). We illustrate four cases: (A) **Legit Reference:** The reconstruction adheres closely to the valid trajectory. (B) **Wrong Appearance:** The input features unseen shapes (Visual refs D, E, F), causing high reconstruction error. (C) **Random Transitions:** The input uses valid shapes but violates the learned transition graph (e.g., direct jumps between non-adjacent states). (D) **Sharp Transitions:** The input lacks the smooth interpolation characteristic of genuine hologram transitions.

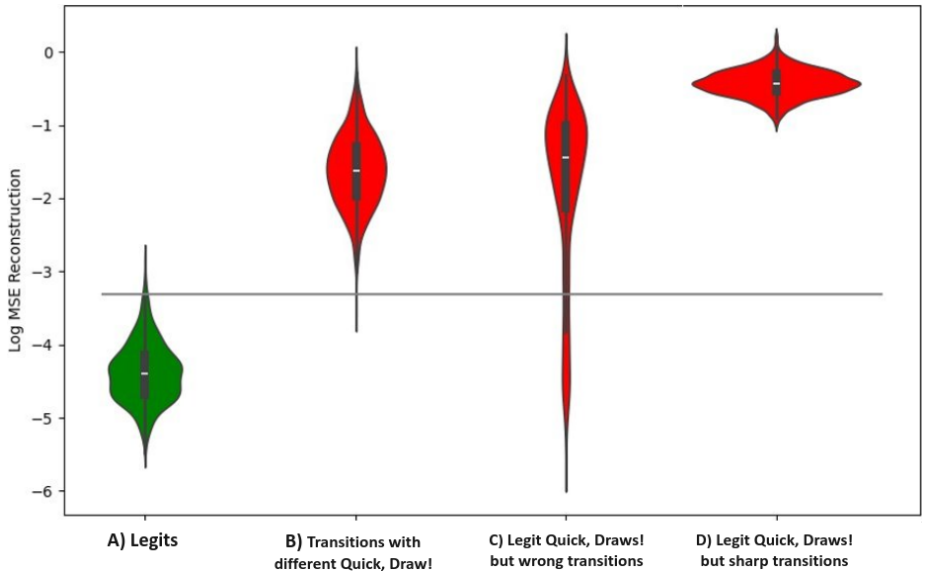

Figure 4.6 – Distribution of the Log-MSE reconstruction error across the synthetic test set. Valid samples (green) generally exhibit significantly lower reconstruction errors compared to fraudulent samples (red), which include wrong appearances, invalid transition paths, and sharp discontinuities. The horizontal dashed line represents the decision threshold τ, calibrated at the 99th percentile of the legitimate distribution (corresponding to a fixed False Positive Rate (FPR) of 1%).

the recall (True Positive Rate) for each fraud category in Table 4.3. Additionally, Figure 4.6 illustrates the separation between the error distributions of legitimate and fraudulent samples.

We achieved 100% recall on both wrong appearances and sharp transitions. The lower recall (85%) on random transitions is intrinsically expected, as stochastic generation occasionally yields valid paths by chance. This preliminary evaluation confirms that masked sequence modeling effectively captures transition validity, enabling robust detection of various temporal anomalies without requiring access to fraudulent samples during training.

Qualitative Analysis. To better understand the model's behavior, we visualize the trajectories of frame embeddings in the projection space. Figure 4.5 displays the input trajectories alongside their reconstructed counterparts (after masking), for the following cases: Legit "3 phases" and fraud regimes including "wrong Quick, Draw!" "wrong transitions," and "sharp transitions," illustrating the model's response to different temporal anomalies.

We observe that for Legit sequences, the reconstructed embeddings closely follow the input trajectory, indicating accurate modeling of valid dynamics. In contrast, fraudulent sequences exhibit larger deviations in the reconstructed embeddings, particularly around anomalous transitions or appearances, leading to

elevated reconstruction errors, and "jumps" in the embedding space. This provides preliminary support for the hypothesis that masked sequence modeling can effectively capture temporal regularities and identify deviations corresponding to various fraud types. It also highlights the potential of this approach to analyze *coverage* (variety) of legitimate dynamics (in complement of their *conformity*), and possibly even enable few-shot matching against reference sequences if the projector is sufficiently general. However, care must be taken to ensure the projector does not inadvertently map attack samples onto the legitimate manifold, especially when trained solely on legitimate data.

4.2.4 Discussion and Limitations

This section explored extending the WSL approach to sequence modeling specifically to capture *transition validity*, a critical dimension absent in frame-wise verification. While validation was restricted to synthetic data, the results demonstrate the approach's promise in detecting various temporal anomalies. The use of a controlled synthetic dataset—featuring distinct OVD appearances and specific fraud scenarios—confirmed that masked sequence prediction in embedding space is effective against gross anomalies. Crucially, these findings demonstrate that *transition validity* is a learnable signal, capable of being modeled without access to real fraudulent samples.

The explicit construction of a structured embedding space enables several directions for future research. Beyond simple anomaly detection, such a space could facilitate few-shot matching against reference sequences or provide a metric for the *coverage* (variety) of legitimate dynamics observed during a session. The proposed smoothness regularization appears to offer a practical remedy to the representational collapse observed when training on limited legitimate data. However, ensuring robust separation of visually distinct OVD states—and preventing the mapping of sophisticated attack samples onto the legitimate manifold—requires further investigation. Future work should explore strengthening the projector, possibly by leveraging foundation models (adapted to focus on fine-grained holographic content), incorporating physical constraints like angular consistency, or diversifying the training set with multiple OVD models to enforce a more discriminative latent structure. Integrating contrastive learning objectives on groups of frames to encourage diverse yet valid dynamics is another potential avenue.

4.3 Conclusion

This chapter presented two exploratory approaches for transition-aware OVD verification, i.e., assessing the validity of temporal dynamics across video sequences, which cannot be assessed from isolated frames alone.

We proposed and evaluated two sequence-level methods—HoloVerif-Span and Masked Sequence Modeling (MSM)—both specifically designed to train exclusively on Legit video sequences, aligning with practical deployment scenarios where real fraud samples are scarce or unavailable.

The **HoloVerif-Span** method leverages a VideoMAE backbone to classify short spans of frames after hologram-focused preprocessing. By generating non-Valid training samples through temporal corruptions (frame repetition, shuffling, segment reordering) and sampling from non-OVD regions, the classifier learns to distinguish valid holographic dynamics from abnormal patterns without ever seeing real attacks during training. However, detection rates remain lower on sophisticated dynamic attacks that closely mimic genuine OVD behavior. Furthermore, while this approach enable the verification of transition validity (dynamic appearance conformity), it does not directly address the coverage of legitimate dynamics.

The **Masked Sequence Modeling (MSM)** approach takes an unsupervised anomaly detection perspective: a frame encoder with smoothness regularization produces structured embeddings, and an LSTM learns to reconstruct masked portions of embedding trajectories. Validation on synthetic data confirmed the core method, enabling threshold-based detection for out of detection. The main limitations of this approach lie 1) in its projection stage as the learned projector must both focus on holographic content and separate distinct appearances—which is challenging when trained solely on legitimate data—and 2) in the absence of appearence coverage modeling at the current stage, despite encouraging preliminary results.

The design of these rather different approaches enabled us to explore distinct paradigms. HoloVerif-Span provides end-to-end supervised learning with strong results on no-OVD and lighting-based attacks, yet requires careful design of synthetic negatives. MSM offers principled unsupervised detection grounded in reconstruction error, yet has only been validated on synthetic data. Table 4.4 summarizes the strengths and limitations of each approach.

Several limitations of this exploratory where identified. First, the two methods have not been evaluated under the same rigorous protocol as Section 3.5.2, and direct comparison with frame-level approaches remains incomplete. Second, Masked Sequence Modeling has only been validated on synthetic data; extension to MIDV-Holo and MIDV-DynAttack is a necessary next step.

These limitations point to clear research directions. **Harder negative mining**—using generative models (GANs, diffusion) to synthesize near-boundary attacks—could populate the gray zone between valid and invalid dynamics, tightening the learned decision boundary and provide a way to address data scarcity. This should improve the performance of methods based on supervised learning, such as HoloVerif-Span. **Enhanced projector training** for MSM is critical to ensure robust separation of distinct holographic appearances, possibly by leveraging foundation models adapted to holographic content or incorporating physical

Table 4.4 – Summary of transition-aware OVD verification methods. Each approach addresses different aspects of the verification problem, with complementary strengths and limitations.

Method	Description	Strengths	Limitations
HoloVerif-Span (Ours)	VideoMAE-based span classifier with background subtraction and temporal corruption for negative sampling.	Trains without real fraud videos Focuses on holographic content	Limited on dynamic attacks mimicking real OVDs Swap detection remains partial Synthetic negatives may not cover sophisticated forgeries No direct modeling of appearance coverage
Masked Sequence Modeling (Ours)	Frame encoder with smoothness loss + LSTM for masked embedding reconstruction; anomaly detection via reconstruction error.	Unsupervised anomaly detection Captures transition validity Detects wrong appearances and sharp transitions Validated concept on synthetic data	Only tested on synthetic dataset Requires extension to real OVD data Projector must separate distinct appearances No direct modeling of appearance coverage (despite potential)

constraints like angular consistency. Integrating contrastive learning objectives on groups of frames to encourage diverse yet valid dynamics is another potential avenue. Incorporating **appearance coverage modeling** into the MSM framework is another promising direction, but practical strategies remain to be explored. More generally, **extensive validation on real OVD data** is essential to confirm the practical utility of these methods and refine their designs.

CHAPTER 5

CONCLUSION

This thesis addressed the challenge of automatic verification of Optically Variable Devices (OVDs) on identity documents captured with commodity smartphones under realistic, uncontrolled conditions. We operated within the demanding constraints of remote Know Your Customer (KYC) workflows: short video sequences, heterogeneous capture devices, variable illumination, large amount of existing document templates, and—critically—severe scarcity of fraud samples for training. Throughout this work, we pursued a central question: *how can we design verification systems that reliably distinguish legitimate "holographic" content (i.e., OVD behavior) from increasingly sophisticated attacks, while remaining practical for industrial deployment?*

We identified key limitations in classical attack detection paradigms, which rely on supervised learning from known fraud samples. We showed in Chapter 2 that such methods struggle to generalize to unseen attack types, due to the inherent diversity of possible forgeries and the limited availability of representative training data. In order to design a system which can generalize to unseen attack types, we proposed methods that explicitly model the authentic properties of OVDs, and decomposed this problem into complementary sub-tasks: *appearance conformity* (does the captured content match the expected holographic appearance/dynamics?) and *appearance coverage* (does the captured content sufficiently explore the OVD state space?). We further refined *appearance conformity* into *static appearance verification* (frame-level holographic behavior)—which we addressed in Chapter 3—and *dynamic appearance verification* (transition validity over time)—which we explored in Chapter 4.

5.1 Summary of Contributions

We contributed in three complementary axes: evaluation resources, learning-based verification methods, and exploratory temporal modeling.

Datasets and Evaluation Protocol. Recognizing that progress in OVD verification was motivated by the lack of realistic, publicly available benchmarks, we introduced MIDV-DynAttack—a test-only dataset of 1,200 video sequences featuring ten distinct attack types not present in existing public resources. Un-

like prior datasets that focused on static template fraud or simple recaptures, MIDV-DynAttack includes dynamic attacks that actively manipulate lighting, swap templates during acquisition, or employ alternate holographic foils. The accompanying evaluation protocol enforces identity separation and reserves part of attacks types for testing only, effective challenging generalization to unseen fraud families.

Frame-level Verification Methods. We developed and systematically compared three data-driven approaches for OVD verification, each addressing different aspects of the problem:

- **Direct Classifier**: A straightforward binary classifier trained with video-level labels applied directly to individual frames. While conceptually simple, this approach proved surprisingly effective against document swapping attacks, demonstrating that naive data-driven methods can capture useful signals even without explicit hologram modeling.
- **HoloVerif** (HoloVerif): A methodological contribution, combining background subtraction to isolate holographic signals from static document content with pseudo-labeling and targeted augmentations. By explicitly guiding the network to focus on holographic regions, HoloVerif achieved state-of-the-art performance on dynamic attacks and demonstrated strong cross-dataset generalization—even to document families and capture pipelines unseen during training.
- **WSL** (Weakly Supervised Learning): A contrastive approach that learns discriminative embeddings using only video-level labels. By encouraging frames from legitimate sequences to cluster separately from attack sequences, this method captures both *appearance conformity* and *appearance coverage* without requiring frame-wise annotations. Notably, WSL can be trained using legitimate samples alone, making it valuable when fraud data is scarce.

Experimental evaluation on MIDV-Holo and MIDV-DynAttack revealed complementary failure modes: Direct Classifier excels at swap attacks but struggles with dynamic holograms; HoloVerif handles dynamic attacks effectively but is challenged by alternating template fraud. This complementarity suggests that production systems should combine multiple verification paradigms rather than relying on a single approach.

Exploratory Temporal Modeling. Recognizing that frame-level methods cannot enforce temporal consistency—a fundamental property of legitimate OVD dynamics—we explored sequence-level OVD verification via two directions, extending the HoloVerif and WSL.

- **HoloVerif-Span (exploratory).** This variant present early work extending HoloVerif to classify temporal segments extracted from the full video sequence, after applying similar same hologram-centered preprocessing as HoloVerif. Invalid training samples are generated through temporal corruption augmentations—frame repetition, shuffling, segment reordering—and by sampling regions known to lack OVD content. This allows the classifier to learn to distinguish legitimate holographic transitions from anomalous patterns without requiring real attack samples during training.

- **Masked Sequence Modeling (MSM) (exploratory).** This approach uniquely addresses anomaly detection through masked temporal modeling. Using a two-stage pipeline (contrastive frame encoder with smoothness regularization based on the WSL method, followed by sequence-to-sequence model based masked prediction), we demonstrated on synthetic data that reconstruction error can detect temporal anomalies (sharp transitions, wrong OVD transitions) without requiring any fraud samples during training. While this work remains exploratory, it validates the potential of temporal modeling and identifies key requirements (smooth embedding trajectories, efficient recurrent architectures) for future development on real OVD data.

Table 5.1 provides a consolidated overview of all verification methods developed in this thesis, highlighting their use cases, training paradigms, and requirements.

5.2 Critical Analysis of Proposed Methods

5.2.1 Analysis: Strengths, Weaknesses, and Recommendations

Direct Classifier. This naive classifier learns to distinguish Legit from Non-Legit frames using any available signal, including dataset biases (background, lighting, device fingerprints). While this limits interpretability and generalization, it makes the method unexpectedly robust to swap attacks, which introduce distributional shifts detectable by a well-trained classifier. The risk is that learned features may not generalize beyond the training distribution. Despite an apparent appealing performance on existing (and limited) datasets, Direct Classifier should not be used in isolation but as a complementary attack detector.

HoloVerif. This approach eliminates dependency on real attack samples by using synthetic temporal corruptions as negatives—a practical advantage for rapid deployment. However, such corruptions may poorly approximate sophisticated attacks maintaining temporal coherence, and validation remains limited to constrained experimental settings. Furthermore, this approach is not natively able to

Table 5.1 – Summary of proposed OVD verification methods. "Specific" refers to verifying a known OVD model; "Generic" refers to detecting any holographic content. "Attack samples" indicates whether fraud data is required for training or threshold calibration.

Method	**Use Case(s)**	**Method Type**	**Training Paradigm**	**Attacks for Training?**	**Attacks for Calibration?**
Direct Classifier	Specific verification	Binary classification	Supervised	Yes	Yes
HoloVerif	Specific verification	Classification with preprocessing	Supervised (pseudo-labels)	Optional	Yes
WSL	Generic detection; Specific (partial)	Contrastive embedding	Weakly supervised	Optional	Yes
HoloVerif-Span (exploratory)	Specific verification (with transition validity)	Classification with preprocessing	Supervised (pseudo-labels & synthetic frauds)	Optional	Optional
MSM (exploratory)	Specific verification (with transition validity)	Masked sequence reconstruction	Unsupervised (Legit-only)	Optional	Optional

verify *appearence coverage*—whether the captured sequence sufficiently explores the OVD state space. HoloVerif-Span was designed to overcome some of these limitations within fraud-scarce scenarios but would require more work and the usage of real world frauds to be fully usable.

WSL. The contrastive embedding approach offers flexibility: it can be trained with Legit samples alone, learns a meaningful latent space, and generalizes reasonably to unseen OVD types. However, its decision mechanism (mean pairwise distance) captures only *coverage* (visual diversity) rather than *conformity* (correct appearance). Consequently, it fails against attacks that produce high visual variance—such as dynamic holographic foils—since these are indistinguishable from legitimate coverage. WSL is best suited as a first-line filter or for scenarios where fraud data is entirely unavailable. Extended training data may enable this method to better capture *appearance conformity*, but this remains to be explored.

HoloVerif-Span (exploratory). This variant present early work extending our leading HoloVerif method to classify temporal segments extracted from the full video sequence, after applying similar hologram-centered preprocessing as HoloVerif. The improved generation of Non-Valid samples from Valid ones enable to train the model using Legit videos only. However, while preliminary results showed that robustness to dymamic attacks may be improved, the lack of native *appearance coverage* verification limits the potential of this approach. Furthermore, it is still very vulnerable to the photo replacement attack, revealing the need for a complementary *spatial coverage* verification.

Masked Sequence Modeling (MSM) (exploratory). This approach uniquely addresses *transition validity*—whether successive frames follow plausible temporal trajectories. Trained only on Legit Synthetic sequences, it can detect sharp or invalid transitions without exposure to fraud samples. Current limitations include sensitivity to projector quality, lack of *appearance coverage* implementation (though promising potential), and a potential vulnerability to photo replacement as well, in the absence of *spatial coverage* verication. Overall, the method is promising but requires validation on real OVD data.

5.2.2 Recommended Deployment Strategy

Based on our experimental findings, we recommend a **layered verification architecture** that combines complementary methods:

1. **First layer: Known attack detection.** Apply Direct Classifier as a complementary attack detector to catch printed documents and other distributional anomalies that evade appearance-based verification.

2. **Second layer: Generic OVD detection.** Apply WSL to filter sequences with no detectable holographic behavior. This catches trivial attacks (static printouts, screen recaptures without holographic overlays) with minimal computational cost.
3. **Third layer: Specific OVD verification.** Apply HoloVerif to verify appearance conformity against the expected OVD model. This catches dynamic attacks that pass the previous layers by employing alternate holographic content. Further more verifies temporal consistency.
4. **Layer 4 (future): Temporal and spatial verification.** Deploy sequence-level methods (HoloVerif-Span, MSM) to enforce transition validity and verify that captured sequences sufficiently cover the OVD manifold—both temporally (plausible state transitions) and spatially (holographic regions across the document).
5. **Layer 5 (out of the scope of this work): User behavior and contextual analysis.** Extend the verification scope beyond isolated video sessions to incorporate session-level and cross-session signals. This layer would monitor user behavior patterns for anomalies by aggregating confidence indicators across the full enrollment flow. Such contextual fraud detection would leverage all available data, from device sensor and historical user activity to flag suspicious behaviors that may not be detectable from the document capture alone.

This layered approach leverages the complementary failure modes observed in our experiments: no single method succeeds across all attack types, but their combination provides broader coverage.

5.3 Limitations

Despite the progress achieved, several fundamental limitations constrain the current state of OVD verification research.

Data Scarcity and Bias. All data-driven methods are ultimately limited by the training distribution. Current public datasets feature only a handful of OVD models, captured with limited device diversity, under restricted conditions. Models trained on this data may learn spurious correlations—device fingerprints, background patterns, lighting artifacts—rather than true holographic features. The risk of capturing dataset biases rather than generalizable representations is particularly acute for attack detection methods like Direct Classifier.

Single-ovd Evaluation. Both MIDV-Holo and MIDV-DynAttack feature only one type of OVD. While cross-dataset experiments on MIDV-2020 and internal data suggest some generalization, rigorous multi-OVD evaluation remains infeasible with current resources. Methods that perform well on one hologram type

may fail on others with different visual characteristics, state spaces, or transition dynamics. Furthermore, OVDs deployed on genuine identity documents—such as advanced *kinegrams*, color-shifting elements, or multi-layer diffractive structures—typically exhibit substantially greater complexity than current public datasets can represent.

Frame-level Processing. The methods presented in Chapter 3 treat videos as unordered sets of frames, discarding temporal structure. This makes them fundamentally vulnerable to attacks that present valid individual frames in implausible sequences (e.g., document swapping, template alternation). While Chapter 4 explored temporal modeling, this work remains preliminary and has not been validated on real OVD data.

Attack Detection Paradigm. Methods like Direct Classifier learn to detect known fraud patterns rather than verify authentic properties. This paradigm is inherently fragile: it generalizes only to attacks similar to those seen during training. Novel attack vectors that differ substantially from the training distribution may evade detection entirely. This type of verification must be used along Model Verification approaches as proposed by HoloVerif and some temporal anomaly approaches.

5.4 Perspectives and Future Work

The limitations identified above point to several promising research directions.

5.4.1 Data Resources

Larger and More Diverse Datasets. Progress in OVD verification is fundamentally limited by data availability. Future efforts should prioritize:

- **Multi-OVD benchmarks:** Datasets featuring diverse hologram types (*kinegrams*, color-shifting inks, micro-optics) from different document families and issuers.
- **Device and identity separation:** Test sets with devices and identities strictly disjoint from training, preventing exploitation of acquisition biases.
- **Realistic attack diversity:** Broader coverage of emerging attack vectors, including deepfake-enhanced documents and adversarial perturbations.

Synthetic Data Generation. Given the practical and legal difficulties of collecting authentic fraud samples, synthetic generation offers a compelling research

direction. Generative models (Generative Adversarial Networks (GANs), diffusion models) and computer generated graphics could address data scarcity at multiple levels:

- **Frame-level synthesis:** Generating realistic attack samples to expose classifiers to diverse forgery patterns, and augmenting legitimate frames with controlled variations to improve robustness.
- **Sequence-level synthesis:** Creating complete video sequences with specified temporal properties—both valid sequences exhibiting diverse legitimate transitions, and invalid sequences containing specific anomaly types (wrong transitions, sharp discontinuities, implausible state orderings).
- **Hard negative mining:** Synthesizing challenging boundary cases for contrastive training, improving discrimination at decision boundaries where current methods struggle.

Preliminary experiments with synthetic temporal data in Chapter 4 suggest this direction is viable; scaling to realistic holographic content generation represents a natural next step.

5.4.2 Pipeline Improvements

Temporal Modeling. The exploratory work in Chapter 4 should be extended to real OVD data:

- **Validation on real data:** Evaluating the proposed temporal methods (HoloVerif-Span, MSM) on MIDV-Holo sequences.
- **Transition graph learning:** Explicitly modeling which OVD visual states can follow which, enabling detection of impossible or improbable state sequences.
- **Group-wise contrastive learning:** Extending the contrastive paradigm from individual frames to *groups of consecutive frames*, capturing both appearance and transition properties in a unified embedding space.

High-quality Reference Acquisitions. In this manuscript we studied uncontrolled condition but industrial settings may permit high-quality reference captures (controlled lighting, calibrated devices) for some OVD models. Further raising questions on how to leverage such references to improve verification:

- **Training-time benefits:** Exploiting controlled acquisitions to obtain richer signals—such as precise illumination angles, tilt speeds, and ground-truth OVD annotations—that are difficult to extract from unconstrained captures

and could be incorporated as an accurate regularization of contrastive representation learning.

- **Domain gap mitigation:** Developing domain adaptation techniques to bridge the quality gap between high-fidelity references and noisy operational data, ensuring models trained or calibrated on controlled captures generalize to real-world conditions.

Foundation Models for Document Security. Large pretrained models have transformed many vision tasks; adapting this paradigm to OVD verification could address data scarcity through transfer learning. Our first test shows encouraging results on DINO models. Promising directions include:

- **Self-supervised pretraining:** Learning holographic representations from unlabeled document videos at scale—probably assumed to be authentic.
- **Targeted finetuning:** Developing strategies to finetune pretrained encoders so they focus on holographic content rather than background or facial features, potentially through attention guidance or feature selection mechanisms.
- **Video foundation models:** Leveraging pretrained spatiotemporal encoders for temporal anomaly detection, as explored in Section 4.2.

5.5 Closing Remarks

Automatic verification of optically variable devices on identity documents presents a fascinating intersection of computer vision, security, and practical engineering. The visual richness of holographic content—designed over decades to resist human-visible counterfeiting—poses unique challenges for machine learning systems that must generalize from limited data to detect sophisticated, previously unseen attacks.

This thesis has advanced the state of the art in several respects: we introduced the first public benchmark specifically designed for dynamic attack evaluation; we developed methods that achieve strong performance on this challenging benchmark; and we laid the groundwork for temporal modeling approaches that address fundamental limitations of frame-level verification. Yet much remains to be done. The gap between academic research and industrial deployment remains substantial, and the arms race between attackers and defenders shows no signs of abating.

We hope that the datasets, methods, and insights presented here will serve as a foundation for future research in this critical domain. As identity verification

becomes ever more central to digital life, the security of the documents that anchor these identities will only grow in importance.

BIBLIOGRAPHY

An, Jinwon and Sungzoon Cho (2015). "Variational autoencoder based anomaly detection using reconstruction probability". In: *Special lecture on IE* 2.1, pp. 1–18 (cit. on p. 74).

Arlazarov, Vladimir V., Konstantin Bulatov, Timofey Chernov, and Vladimir L. Arlazarov (2019). "MIDV-500: A Dataset for Identity Documents Analysis and Recognition on Mobile Devices in Video Stream". In: *Computer Optics* 43.5 (cit. on pp. 16, 17).

Ay, Betul (2022). "Open-Set Learning-Based Hologram Verification System Using Generative Adversarial Networks". In: *IEEE Access* 10, pp. 25114–25124 (cit. on p. 34).

Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun (2018). *Convolutional Sequence Modeling Revisited*. URL: https://openreview.net/forum?id=rk8wKk-R- (cit. on p. 73).

Balntas, Vassileios, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk (2016). "Learning local feature descriptors with triplets and shallow convolutional neural networks". In: *Procedings of the British Machine Vision Conference 2016*. York, UK: British Machine Vision Association, pp. 119.1–119.11 (cit. on p. 44).

Boned, Carlos, Maxime Talarmain, Nabil Ghanmi, Guillaume Chiron, Sanket Biswas, Ahmad Montaser Awal, and Oriol Ramos Terrades (2024). "Synthetic dataset of id and travel documents". In: *Scientific Data* 11.1, p. 1356 (cit. on pp. 16, 18).

Bulatov, Konstantin, Ekaterina Emelianova, Daniil Tropin, Natalya Skoryukina, Yulia Chernyshova, Alexander Sheshkus, Sergey Usilin, Zuheng Ming, Jean-Christophe Burie, Muhammad Muzzamil Luqman, and Vladimir V. Arlazarov (2022). "MIDV-2020: A Comprehensive Benchmark Dataset for Identity Document Analysis". In: *Computer Optics* 46.2 (cit. on pp. 16, 17, 55).

Bulatov, Konstantin, Daniil Matalov, and Vladimir V. Arlazarov (2020). "MIDV-2019: Challenges of the modern mobile-based document OCR". In: *Proceedings of the twelfth International Conference on Machine Vision (ICMV 2019)*, p. 64 (cit. on pp. 16, 17).

Burie, JC., J. Chazalon, M. Coustaty, S. Eskenazi, M.M. Luqman, M. Mehri, N. Nayef, JM. Ogier, S. Prum, and M. Rusiñol (2015). "ICDAR2015 competition on smartphone document capture and OCR (SmartDoc)". In: *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1161–1165 (cit. on p. 17).

Chapel, Marie-Neige, Musab Al-Ghadi, and Jean-Christophe Burie (2023). "Authentication of Holograms with Mixed Patterns by Direct LBP Comparison". In: *Proceedings of the 25th International Workshop on Multimedia Signal Processing*, pp. 1–6 (cit. on p. 34).

Chen, Yuhang, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang (2023). *ImDiffusion: Imputed Diffusion Models for Multivariate Time Series Anomaly Detection*. arXiv: 2307.00754 [cs.LG]. URL: https://arxiv.org/abs/2307.00754 (cit. on p. 73).

Chernyshova, Yulia, Ekaterina Emelianova, Alexander Sheshkus, and Vladimir Arlazarov (2021). "MIDV-LAIT: A Challenging Dataset for Recognition of IDs with Perso-Arabic, Thai, and Indian Scripts". In: *Proceedings of the 16th International Conference Document Analysis and Recognition*. Pp. 258–272 (cit. on pp. 16, 17).

Chung, Junyoung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio (2014). "Empirical evaluation of gated recurrent neural networks on sequence modeling". In: *arXiv preprint arXiv:1412.3555* (cit. on p. 72).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255 (cit. on p. 56).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186 (cit. on p. 73).

European Union, Council of the (2022). *PRADO - Public Register of Authentic travel and identity Documents Online, v. 12344/22*. URL: https://www.consilium.europa.eu/prado/en/prado-glossary.html (visited on 01/25/2024) (cit. on p. 5).

Fu, Yiwei and Feng Xue (July 2022). "MAD: Self-Supervised Masked Anomaly Detection Task for Multivariate Time Series". In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. URL: http://dx.doi.org/10.1109/IJCNN55064.2022.9892218 (cit. on p. 74).

Al-Ghadi, Musab, Zuheng Ming, Petra Gomez-Krämer, Jean-Christophe Burie, Mickaël Coustaty, and Nicolas Sidere (2023). "Guilloche Detection for ID Authentication: A Dataset and Baselines". In: *Proceedings of the 25th International Workshop on Multimedia Signal Processing*, pp. 1–6 (cit. on pp. 16, 18).

Goyal, Raghav, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic (Oct. 2017). "The "Something Something" Video Database for Learning and Evaluating Visual Common Sense". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (cit. on p. 67).

Graves, Alex and Jürgen Schmidhuber (2005). "Framewise phoneme classification with bidirectional LSTM and other neural network architectures". In: *Neural networks* 18.5-6, pp. 602–610 (cit. on p. 73).

Guan, Hong, Yancheng Wang, Lulu Xie, Soham Nag, Rajeev Goel, Niranjan Erappa Narayana Swamy, Yingzhen Yang, Chaowei Xiao, Jonathan Prisby, Ross

Maciejewski, and Jia Zou (2024). *IDNet: A Novel Dataset for Identity Document Analysis and Fraud Detection*. eprint: arxiv/2408.01690 (cit. on pp. 16, 18).

Hasan, Mahmudul, Jonghyun Choi, Jan Neumann, Amit K Roy-Chowdhury, and Larry S Davis (2016). "Learning temporal regularity in video sequences". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 733–742 (cit. on p. 74).

He, Kaiming, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick (2022). "Masked autoencoders are scalable vision learners". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009 (cit. on p. 73).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778 (cit. on p. 56).

Hinton, G. E. and R. R. Salakhutdinov (2006). "Reducing the Dimensionality of Data with Neural Networks". In: *Science* 313.5786, pp. 504–507 (cit. on p. 73).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780 (cit. on p. 72).

Hoogeboom, Emiel, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans (2022). "Autoregressive Diffusion Models". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Lm8T39vLDTE (cit. on p. 73).

Howard, Andrew, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le (2019). "Searching for MobileNetV3". In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324 (cit. on p. 56).

Iqbal, Hasan, Umar Khalid, Chen Chen, and Jing Hua (2024). "Unsupervised Anomaly Detection in Medical Images Using Masked Diffusion Model". In: *Machine Learning in Medical Imaging*. Ed. by Xiaohuan Cao, Xuanang Xu, Islem Rekik, Zhiming Cui, and Xi Ouyang. Cham: Springer Nature Switzerland, pp. 372–381 (cit. on p. 73).

Jongejan, J, H Rowley, T Kawashima, J Kim, and N Fox-Gieg (2017). *The quick, draw! dataset* (cit. on pp. 66, 79).

Kada, Oumayma, Camille Kurtz, Cuong van Kieu, and Nicole Vincent (2022). "Hologram Detection for Identity Document Authentication". In: *Proceedings of the 3rd International Conference on Pattern Recognition and Artificial Intelligence*, pp. 346–357 (cit. on p. 34).

Kingma, Diederik P and Max Welling (2013). "Auto-Encoding Variational Bayes". In: *arXiv preprint arXiv:1312.6114* (cit. on p. 73).

Kokhlikyan, Narine, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson (2020). *Captum: A unified and generic model interpretability library for PyTorch*. arXiv: 2009.07896 [cs.LG] (cit. on p. 47).

Koliaskina, L. I., E. V. Emelianova, D. V. Tropin, V. V. Popov, K. B. Bulatov, D. P. Nikolaev, and V. V. Arlazarov (2023). "MIDV-Holo: A Dataset for ID Document Hologram Detection in a Video Stream". In: *Proceedings of the 17th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 486–503 (cit. on pp. 16, 19, 20, 28, 35–37, 43, 49, 62).

Korshunov, Pavel, Amir Mohammadi, Vidit Vidit, Christophe Ecabert, and Sébastien Marcel (2025a). "FantasyID: A dataset for detecting digital manipulations of ID-documents". In: *Proceedings of the IEEE International Joint Conference on Biometrics* (cit. on pp. 16, 18).

Korshunov, Pavel, Vidit Vidit, Amir Mohammadi, Christophe Ecabert, Nevena Shamoska, Sébastien Marcel, Zeqin Yu, Ye Tian, Jiangqun Ni, Lazar Lazarevic, Renat Khizbullin, Anastasiia Evteeva, Alexey Tochin, Aleksei Grishin, Anjith George, Daniel Dealcala, Tamas Endrei, Javier Muñoz-Haro, Ruben Tolosana, Ruben Vera-Rodriguez, Aythami Morales, Julian Fierrez, György Cserey, Hardik Sharma, Sachin Chaudhary, Akshay Dudhane, Praful Hambarde, Amit Shukla, Prateek Shaily, Jayant Kumar, Ajinkya Hase, Satish Maurya, Mridul Sharma, and Pallav Dwivedi (Oct. 2025b). "DeepID Challenge of Detecting Synthetic Manipulations in ID Documents". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 510–519 (cit. on p. 18).

Learned-Miller, Erik, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua (2016). "Labeled faces in the wild: A survey". In: *Advances in face detection and facial image analysis*. Springer, pp. 189–248 (cit. on p. 80).

Lerouge, Julien, Guillaume Betmont, Thomas Bres, Evgeny Stepankevich, and Alexis Bergès (2024). *DocXPand-25k: a large and diverse benchmark dataset for identity documents analysis*. eprint: arXiv/2408.01690 (cit. on pp. 16, 17).

Loshchilov, Ilya and Frank Hutter (Jan. 4, 2019). *Decoupled Weight Decay Regularization*. arXiv: 1711.05101 (cit. on p. 44).

Malhotra, P (2016). "LSTM-based Encoder-Decoder for Multi-sensor Anomaly Detection". In: *arXiv preprint arXiv:1607.00148* (cit. on p. 74).

Mehta, Sachin and Mohammad Rastegari (2021). "MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer". In: *International Conference on Learning Representations*. arXiv: 2110.02178 (cit. on p. 56).

Neves Junior, Ricardo Batista das, Byron Leite Dantas Bezerra, and Cleber Zanchettin (2024). "DocLightDetect: A New Algorithm for Occlusion Classification in Identification Documents". In: *Proceedings of the 16th IAPR International Workshop on Document Analysis Systems (DAS)*, pp. 196–210 (cit. on pp. 16, 17).

Nicolaou, Anguelos, Vincent Christlein, Edgar Riba, Jian Shi, Georg Vogeler, and Mathias Seuret (2022). "TorMentor: Deterministic dynamic-path, data augmentations with fractals". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2707–2711 (cit. on p. 66).

Oquab, Maxime, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li,

Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski (2023). *DINOv2: Learning Robust Visual Features without Supervision* (cit. on pp. 47, 48, 75).

Park, Eun-Ju, Seung-Yeon Back, Jeongho Kim, and Simon S. Woo (2023). "KID34K: A Dataset for Online Identity Card Fraud Detection". In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5381–5385 (cit. on pp. 16, 18, 26).

Polevoy, Dmitry V., Irina V. Sigareva, Daria M. Ershova, Vladimir V. Arlazarov, Dmitry P. Nikolaev, Zuheng Ming, Muhammad Muzzamil Luqman, and Jean-Christophe Burie (2022). "Document Liveness Challenge Dataset (DLC-2021)". In: *Journal of Imaging* 8.7, p. 181 (cit. on pp. 16, 18, 26).

Pouliquen, Glen, Joseph Chazalon, Guillaume Chiron, Thierry Géraud, and Ahmad Montaser Awal (2025a). "Verification of Dynamic Holographic Behavior in Identity Documents". In: *Proceedings of the 19th International Conference on Document Analysis and Recognition (ICDAR)*. Springer, pp. 323–339 (cit. on pp. 13, 16, 20, 39).

Pouliquen, Glen, Joseph Chazalon, Guillaume Chiron, Thierry Géraud, and Ahmad Montaser Awal (2025b). *MIDV-DynAttack*. International Conference on Document Analysis and Recognition (ICDAR). Zenodo. URL: https://doi.org/10.5281/zenodo.17079529 (cit. on p. 13).

Pouliquen, Glen, Guillaume Chiron, Joseph Chazalon, Thierry Géraud, and Ahmad Montaser Awal (2024). "Weakly Supervised Training for Hologram Verification in Identity Documents". In: *Proceedings of the 18th International Conference on Document Analysis and Recognition (ICDAR)*, pp. 17–33 (cit. on pp. 12, 17).

Riba, Edgar, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski (2020). "Kornia: an open source differentiable computer vision library for pytorch". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3674–3683 (cit. on p. 66).

Siméoni, Oriane, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. (2025). "Dinov3". In: *arXiv preprint arXiv:2508.10104* (cit. on p. 56).

Soares, Álysson, Ricardo das Neves Junior, and Byron Bezerra (2020). "BID Dataset: a challenge dataset for document processing tasks". In: *Proceedings of the XXXIII Conference on Graphics, Patterns and Images*, pp. 143–146 (cit. on pp. 16, 17).

Soukup, Daniel and Reinhold Huber-Mörk (2017). "Mobile hologram verification with deep learning". In: *IPSJ Transactions on Computer Vision and Applications* 9 (cit. on p. 35).

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *Proceedings of the 34th International Conference on*

Machine Learning - Volume 70. ICML'17. Sydney, NSW, Australia: JMLR.org, pp. 3319–3328. arXiv: 1703.01365 (cit. on p. 47).

Tong, Zhan, Yibing Song, Jue Wang, and Limin Wang (2022). "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training". In: *Advances in neural information processing systems* 35, pp. 10078–10093 (cit. on pp. 64, 67, 73).

Uniting Aviation: A United Nations Specialized Agency (Sept. 2022). URL: https://www.icao.int/publications/pages/publication.aspx?docnum=9303 (cit. on p. 20).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30 (cit. on p. 72).

Wang, Zhou, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612 (cit. on p. 65).

Wightman, Ross (2019). *PyTorch Image Models*. https://github.com/rwightman/pytorch-image-models (cit. on p. 56).