THÈSE PRÉSENTÉE

POUR OBTENIR LE GRADE DE

DOCTEUR
DE L'UNIVERSITÉ DE BORDEAUX

ECOLE DOCTORALE MATHÉMATIQUES ET

INFORMATIQUE

SPÉCIALITÉ : INFORMATIQUE

Par **Arthur LONGUEFOSSE**

APPRENTISSAGE PROFOND POUR LA CONVERSION D'IRM VERS TDM EN IMAGERIE THORACIQUE

Deep learning for MR to CT translation in thoracic imaging

Soutenue le 18 décembre 2024

Membres du jury :

Mme. Jenny BENOIS-PINEAU	Professeure	Univ. de Bordeaux	Présidente
M. Serge MIGUET	Professeur	Univ. Lumière Lyon 2	Rapporteur
M. José Maria MARTINEZ SANCHEZ	Professeur	Univ. de Madrid	Rapporteur
M. Nicolas NORMAND	Professeur	Univ. de Nantes	Examinateur
M. Pascal DESBARATS	Professeur	Univ. de Bordeaux	Directeur
M. Fabien BALDACCI	Maître de Conférences	Univ. de Bordeaux	Co-directeur
M. Gaël DOURNES	Professeur	CHU de Bordeaux	Invité
M. Baudouin DENIS DE SENNEVILLE	Directeur de Recherche	Univ. de Bordeaux	Invité

Apprentissage profond pour la conversion d'IRM vers TDM en imagerie thoracique

Résumé : L'imagerie thoracique présente des défis importants, chaque modalité d'imagerie ayant ses propres limitations. Le scanner (CT), référence en imagerie pulmonaire, offre une haute résolution spatiale, mais utilise des rayonnements ionisants, posant des risques pour les patients nécessitant des examens fréquents. À l'inverse, l'IRM pulmonaire offre une alternative sans radiation, mais est limitée par des problèmes techniques tels qu'un faible contraste et des artefacts, limitant son adoption clinique à grande échelle. Récemment, l'IRM à temps d'écho ultracourt (UTE-MRI) a montré un potentiel pour surmonter certaines de ces limitations, mais elle ne parvient toujours pas à atteindre la haute résolution et la qualité d'image du scanner, notamment pour l'évaluation détaillée des structures.
L'objectif principal de cette thèse est de développer et valider des modèles basés sur l'apprentissage profond pour synthétiser des images de scanner TDM à partir d'IRM UTE. Plus précisément, nous visons à évaluer la qualité des images, la reconstruction anatomique et l'applicabilité clinique de ces images de synthèse par rapport aux IRM UTE originales et aux scanners CT correspondants.
Dans un premier temps, nous avons exploré les bases de la synthèse d'images médicales, en établissant le socle pour la conversion de l'IRM vers le scanner. Nous avons mis en œuvre un modèle de GAN 2D basé sur le cadre pix2pixHD, optimisé avec la normalisation SPADE et affiné les techniques de prétraitement telles que le rééchantillonnage et l'enregistrement. L'évaluation clinique par des radiologues experts a montré des résultats prometteurs en comparant les images synthétiques aux scans réels. La synthèse a été ensuite améliorée par l'introduction de la perte perceptuelle, qui a affiné les détails structurels et la qualité visuelle, et par l'intégration de stratégies 2.5D pour équilibrer la synthèse 2D et 3D. De plus, nous avons mis l'accent sur un processus de validation rigoureux utilisant des métriques spécifiques à la tâche, remettant en question les métriques globales traditionnelles basées sur l'intensité, en nous concentrant sur la reconstruction précise des structures anatomiques. Dans la dernière étape, nous avons développé un cadre de synthèse 3D robuste et évolutif en adaptant nnU-Net pour la génération de scanner, accompagné d'une fonction de perte priorisant les caractéristiques anatomiques, permettant une meilleure reconstruction des structures critiques telles que les voies respiratoires et les vaisseaux.
Notre travail met en évidence le potentiel des modèles basés sur l'apprentissage profond pour générer des images synthétiques de haute qualité de type scanner à partir d'IRM UTE, offrant une amélioration significative de l'imagerie pulmonaire non invasive. Ces avancées pourraient considérablement améliorer l'applicabilité clinique de l'IRM UTE, en offrant une alternative plus sûre au scanner pour le suivi des maladies pulmonaires chroniques. De plus, un brevet est actuellement en préparation pour l'adoption de notre méthode, ouvrant la voie à une utilisation potentielle en milieu clinique.

Mots-clés : IRM pulmonaire, Apprentissage profond, Conversion inter-modalité, Préservation des structures fines

Laboratoire Bordelais de Recherche en Informatique (LaBRI)
Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

Deep learning for MR to CT translation in thoracic imaging

Abstract: Thoracic imaging faces significant challenges, with each imaging modality presenting its own limitations. CT, the gold standard for lung imaging, delivers high spatial resolution but relies on ionizing radiation, posing risks for patients requiring frequent scans. Conversely, lung MRI, offers a radiation-free alternative but is hindered by technical issues such as low contrast and artifacts, limiting its broader clinical use. Recently, UTE-MRI shows promise in addressing some of these limitations, but still lacks the high resolution and image quality of CT, particularly for detailed structural assessment. The primary objective of this thesis is to develop and validate deep learning-based models for synthesizing CT-like images from UTE-MRI. Specifically, we aim to assess the image quality, anatomical accuracy, and clinical applicability of these synthetic CT images in comparison to the original UTE-MRI and real CT scans in thoracic imaging.
Initially, we explored the fundamentals of medical image synthesis, establishing the groundwork for MR to CT translation. We implemented a 2D GAN model based on the pix2pixHD framework, optimizing it using SPADE normalization and refining preprocessing techniques such as resampling and registration. Clinical evaluation with expert radiologists showed promising results in comparing synthetic images to real CT scans. Synthesis was further enhanced by introducing perceptual loss, which improved structural details and visual quality, and incorporated 2.5D strategies to balance between 2D and 3D synthesis. Additionally, we emphasized a rigorous validation process using task-specific metrics, challenging traditional intensity-based and global metrics by focusing on the accurate reconstruction of anatomical structures. In the final stage, we developed a robust and scalable 3D synthesis framework by adapting nnU-Net for CT generation, along with an anatomical feature-prioritized loss function, enabling superior reconstruction of critical structures such as airways and vessels. Our work highlights the potential of deep learning-based models for generating high-quality synthetic CT images from UTE-MRI, offering a significant improvement in non-invasive lung imaging. These advances could greatly enhance the clinical applicability of UTE-MRI, providing a safer alternative to CT for the follow-up of chronic lung diseases. Furthermore, a patent is currently in preparation for the adoption of our method, paving the way for potential clinical use.

Keywords: Lung MRI, Deep Learning, Cross-modality Translation, Fine-structure Preservation

Laboratoire Bordelais de Recherche en Informatique (LaBRI)
Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France

Remerciements

Tout d'abord, un grand merci à mes encadrants Pascal et Fabien, pour leur accompagnement amical et toujours dans la bonne humeur. Vos conseils avisés et votre confiance ont été essentiels tout au long de ces années. Un remerciement tout particulier à Baudouin, dont l'investissement et le soutien constants font de lui un encadrant à part entière. Sans ton expertise et ton implication, cette thèse n'aurait certainement pas atteint ce niveau d'aboutissement.
Je tiens également à remercier chaleureusement les radiologues Gaël et Illyes pour leur précieuse collaboration interdisciplinaire. Votre engagement et votre soutien tout au long de ce projet ont été déterminants pour en renforcer la pertinence et la qualité. Un merci sincère également à Imène et Yannis, avec qui j'ai eu le plaisir de partager des discussions enrichissantes lors des réunions partagées.
Je souhaite aussi exprimer toute ma gratitude aux membres du jury pour leur temps, leur investissement et leurs précieux retours. Merci à Jenny Benois-Pineau, présidente du jury, pour avoir dirigé la soutenance avec bienveillance et rigueur. Un grand merci à Serge Miguet, rapporteur, pour ses remarques détaillées sur la thèse et ses conseils lors de la soutenance. Une pensée particulière pour José Maria Martinez Sanchez, également rapporteur, pour son investissement et son retour. Enfin, merci à Nicolas Normand, examinateur, pour son soutien et ses recommandations précieuses qui ont enrichi cette soutenance.
Mes pensées se tournent également vers mes collègues et amis doctorants, qui ont grandement contribué à rendre ces années mémorables. Merci à Clément et à tout le bureau 267 pour les raclettes inoubliables : Paul, Raquel, Zoé, et bien sûr Ludovic, avec qui j'ai eu le plaisir de brûler ces quelques calories au badminton et au squash. Une pensée particulière pour le bureau 256, avec Edern, Gustavo, et nos délicieux repas partagés au RU. Une pensée émue également pour Abel ; ton sourire est éternel. Merci à tous les doctorants avec qui j'ai échangé en conférences, et notamment à Simon, pour les nombreux moments partagés ensemble. Wer ?
Je tiens à exprimer toute ma reconnaissance à ma famille. À mes parents, pour leur soutien et leurs encouragements constants. À mon frère David, pour ses concerts de guitare, et à ma sœur Alice, pour sa bonne humeur contagieuse et les moments festifs partagés. Un immense merci également à ma belle-famille, pour son accueil chaleureux tout au long de ces années.
Un grand merci également à mes amis, Tony, Jonathan, Pierre, Killian, Corentin, Victor, Romain, Paul, Antho, Clément, Théo, Cédric, pour leur présence, leur amitié et tous ces moments de rire en dehors de la thèse. C'est trop bon.
Enfin, ma plus profonde gratitude va à Alexia, pour ton soutien incommensurable, ton amour et ta patience. Merci d'avoir été à mes côtés tout au long de cette aventure, je te dédie une grande part de cette réussite.

Contents

Résumé étendu en français **1**

Publications **6**

Introduction **7**

1 Fundamentals of Medical Image Synthesis **13**
1.1 Introduction to Image Synthesis . . . 13
1.1.1 Traditional Approaches . . . 13
1.1.2 Machine Learning Approaches . . . 17
1.2 Image-to-Image Translation in Medical Imaging . . . 22
1.2.1 Paired Approaches . . . 23
1.2.2 Unpaired Approaches . . . 25
1.3 Evaluation Metrics for Synthesized Images . . . 27
1.3.1 Intensity-based Metrics . . . 27
1.3.2 Distribution-based Metrics . . . 29
1.3.3 Challenges in Evaluating Medical Image Synthesis . . . 30
1.4 Limitations in Thoracic Imaging . . . 32
1.4.1 Novelty and Dataset Limitations . . . 32
1.4.2 Image Quality Challenges . . . 33
1.4.3 Anatomical Complexity of Thoracic Structures . . . 33
Chapter Conclusion . . . 35

2 MR to CT Synthesis using 2D GANs : Feasibility and Preprocessing Optimization **36**
2.1 Feasibility of MR to CT Lung Translation . . . 37
2.1.1 Exploring 2D GAN Architectures . . . 37
2.1.2 Dataset Insights . . . 39
2.1.3 Ensuring Dataset Pairing . . . 39
2.1.4 Results . . . 44
2.1.5 Discussion . . . 46
2.2 Optimizing Preprocessing for Improved Synthesis . . . 47
2.2.1 Enhancing Registration using Lung Masks . . . 47
2.2.2 Preprocessing Strategies . . . 50
2.2.3 Results . . . 52
2.2.4 Discussion . . . 55
2.3 Clinical Assessment . . . 56
2.3.1 Methodology . . . 56
2.3.2 Results . . . 58

2.3.3 Image Quality . . . 58
2.3.4 Correlations of Structural Alterations . . . 58
2.3.5 Discussion . . . 60
Chapter Conclusion . . . 62

3 Enhancing Synthesis with Perceptual Loss and 2.5D Approaches 64
3.1 Rethinking Evaluation Metrics For Medical Image Synthesis . . . 65
3.1.1 Limitations of Conventional Metrics . . . 65
3.1.2 Introducing Task-Specific Metrics for Anatomical Reconstruction . . . 67
3.2 Optimizing Perceptual Loss for Enhanced Synthesis . . . 69
3.2.1 Implementation Details . . . 70
3.2.2 Exploration of Perceptual Loss Parameters . . . 70
3.2.3 Synthesis of Findings . . . 76
3.3 Bridging the Gap Between 2D and 3D Synthesis : Strategies for 2.5D . . . 78
3.3.1 Overview of Existing 2.5D Methods . . . 79
3.3.2 Implementation Details . . . 80
3.3.3 2.5D Multi-Axial Strategies: Fusion of 2D Networks . . . 80
3.3.4 To 3D or not 3D : That is the Convolution . . . 82
3.3.5 Proposed Hybrid 2.5D Approach . . . 83
3.3.6 Computational Efficiency . . . 86
3.3.7 Synthesis of Findings . . . 86
Chapter Conclusion . . . 88

4 Advanced 3D Synthesis with Scalable Architectures and Anatomical Guidance 89
4.1 Robust and Scalable 3D Synthesis with nnU-Net . . . 91
4.1.1 Introduction to nnU-Net Framework . . . 91
4.1.2 Adapting nnU-Net for 3D Synthesis . . . 92
4.1.3 Validation using Public Challenges and Datasets . . . 94
4.1.4 Findings and Limitations in Thoracic Imaging . . . 98
4.2 Anatomical Feature-Prioritized Loss for Enhanced Synthesis . . . 100
4.2.1 Introduction . . . 100
4.2.2 Definition . . . 102
4.2.3 Validation on MR to CT translation . . . 104
4.2.4 Synthesis of Findings . . . 110
Chapter Conclusion . . . 113

Conclusion and Perspectives 114

Résumé étendu en français

L'imagerie thoracique est un domaine fondamental en médecine, jouant un rôle essentiel dans le diagnostic et le suivi des maladies pulmonaires. Le scanner (CT) est considéré comme la référence en imagerie pulmonaire en raison de sa haute résolution spatiale et de sa capacité à visualiser des détails anatomiques fins. Cependant, l'utilisation de rayonnements ionisants lors des examens CT pose des risques significatifs, notamment pour les patients nécessitant des examens fréquents, comme ceux atteints de maladies pulmonaires chroniques. L'exposition répétée aux radiations augmente le risque de développer des complications à long terme, y compris des cancers induits par les radiations.
En réponse à ces préoccupations, l'imagerie par résonance magnétique (IRM) pulmonaire a émergé comme une alternative non invasive et sans radiation. Néanmoins, l'IRM pulmonaire conventionnelle est limitée par un faible contraste et des artefacts liés aux mouvements respiratoires, dus à la faible densité de protons dans les tissus pulmonaires. Ces limitations entravent l'adoption clinique à grande échelle de l'IRM pour l'imagerie thoracique, car elle n'atteint pas la résolution et la qualité d'image nécessaires pour une évaluation détaillée des structures pulmonaires.
Récemment, l'IRM à temps d'écho ultracourt (UTE-MRI) a montré un potentiel pour surmonter certaines de ces limitations. L'UTE-MRI permet de réduire les effets des mouvements respiratoires et d'améliorer le signal provenant des tissus à relaxation rapide, comme le poumon. Cependant, malgré ces avancées, l'UTE-MRI ne parvient toujours pas à atteindre la haute résolution et la qualité d'image du scanner CT, notamment pour l'évaluation détaillée des structures telles que les bronches et les vaisseaux sanguins.
Face à ces défis, une collaboration entre le Centre Hospitalier Universitaire de Bordeaux et le Laboratoire Bordelais de Recherche en Informatique (LaBRI) a été établie pour combiner l'expertise clinique et les techniques computationnelles avancées. L'objectif central est de tirer parti des méthodes d'apprentissage profond pour convertir les scans IRM UTE en images synthétiques de scanner CT. Cette approche vise à combiner la sécurité et la non-invasivité de l'IRM avec la capacité du scanner CT à visualiser en détail les structures anatomiques, offrant ainsi une alternative plus sûre pour l'imagerie pulmonaire.

Dans le Chapitre 1, nous présentons les fondamentaux de la synthèse d'images médicales. Nous commençons par les approches traditionnelles, telles que les méthodes basées sur l'interpolation et la modélisation statistique, qui ont été les premières à permettre la reconstruction d'images manquantes ou la conversion entre modalités. Avec l'essor du machine learning, des modèles plus sophistiqués ont été développés, exploitant les capacités des réseaux de neurones pour apprendre des représentations complexes des données médicales.
L'introduction du deep learning a révolutionné le domaine, en particulier avec les réseaux de neurones convolutifs (CNN) qui ont démontré une efficacité remarquable en classification et segmentation d'images médicales. Les réseaux antagonistes génératifs (GAN),

quant à eux, ont permis des avancées significatives en synthèse d'images, en apprenant à générer des images réalistes à partir de bruit ou d'autres images. Les modèles de diffusion, plus récents, offrent une approche alternative en modélisant le processus de génération d'images comme une série de transformations progressives.
Nous explorons ensuite la conversion d'image à image en imagerie médicale, qui vise à apprendre une correspondance entre deux modalités d'imagerie, comme l'IRM et le CT. Les approches appariées (paired) nécessitent des paires d'images correspondantes pour l'entraînement, tandis que les approches non appariées (unpaired) utilisent des techniques comme les cycles de cohérence pour surmonter l'absence de correspondances exactes. Ces méthodes ont été appliquées avec succès en neuroimagerie et en imagerie du pelvis, mais leur application en imagerie thoracique représente un nouveau défi unique.
Un aspect crucial de la synthèse d'images est l'évaluation de la qualité des images générées. Les métriques d'évaluation traditionnelles incluent les métriques basées sur l'intensité, telles que l'erreur quadratique moyenne (MSE) et le pic du rapport signal sur bruit (PSNR), qui mesurent les différences globales entre les images synthétiques et de référence. Les métriques basées sur la distribution, comme le Frechet Inception Distance (FID), évaluent la similarité statistique entre les distributions des images générées et réelles. Cependant, dans le contexte médical, ces métriques ne capturent pas nécessairement la pertinence clinique ou la précision anatomique des images synthétisées.
L'application de modèles de synthèse en imagerie thoracique est limitée par plusieurs facteurs. Tout d'abord, il existe peu de jeux de données disponibles publiquement, en raison de la nouveauté des méthodes UTE-MRI et des préoccupations liées à la confidentialité des données médicales. De plus, les images UTE-MRI présentent une résolution et une qualité d'image inférieures à celles du CT, avec des artefacts de mouvement et un faible rapport signal sur bruit. L'anatomie complexe du thorax, avec ses structures fines et dynamiques comme les bronches, les vaisseaux sanguins et le parenchyme pulmonaire, rend la tâche de synthèse d'autant plus difficile.
En conclusion, ce premier chapitre établit les bases nécessaires pour comprendre la synthèse d'images, en particulier dans le contexte de l'imagerie médicale et de la conversion de l'IRM vers le CT. Nous avons mis en évidence les avancées réalisées grâce aux architectures profondes comme les CNN et les GAN, tout en soulignant les défis spécifiques à l'imagerie thoracique. Les limitations actuelles des métriques d'évaluation et le besoin de méthodes adaptées aux particularités du poumon justifient une exploration approfondie, que nous poursuivrons dans les chapitres suivants. Nous aborderons notamment la faisabilité de la synthèse IRM vers CT pour le poumon, en développant des modèles capables de surmonter les obstacles identifiés et en évaluant leur pertinence clinique avec des experts radiologues.

Dans le Chapitre 2, nous présentons une étude approfondie sur la faisabilité et l'optimisation de la synthèse d'images pulmonaires de l'IRM vers le scanner CT en utilisant des réseaux antagonistes génératifs (GAN). L'objectif initial était d'explorer et de comparer divers modèles génératifs, supervisés et non supervisés, pour évaluer leur capacité à générer des images CT synthétiques de haute qualité à partir de données d'IRM UTE.
Nous avons commencé par explorer l'utilisation des GAN 2D pour la synthèse d'images pulmonaires, motivés par les défis associés aux méthodes 3D, tels que les coûts élevés en mémoire et la nécessité de disposer d'un vaste ensemble de données. En nous concentrant sur des modèles appariés et non appariés, nous avons constaté que le modèle supervisé pix2pixHD, particulièrement lorsqu'il est amélioré avec la normalisation spatialement adaptative (SPADE), surpassait systématiquement les autres. La méthode SPADE a aidé

à limiter l'apparition de caractéristiques hallucinées par les GAN, produisant des images CT synthétiques plus fidèles aux scans CT réels. Cette approche souligne le potentiel significatif des modèles GAN appariés pour atteindre une synthèse d'images inter-modalités de haute qualité.
Un accent particulier a été mis sur le rôle crucial du prétraitement des données. Nous avons notamment souligné l'importance du rééchantillonnage basé sur la modalité avec la plus haute résolution. En remettant en question les pratiques de pointe, nous avons choisi d'utiliser la résolution supérieure des images CT pour assurer la préservation des détails anatomiques fins, essentiels pour une synthèse d'images précise, en particulier dans les régions pulmonaires. De plus, l'introduction d'une étape de recalage conditionnel utilisant des masques pulmonaires a amélioré l'alignement entre les images IRM et CT, contribuant à de meilleurs résultats de synthèse.
L'évaluation clinique a fourni des preuves solides que les images CT synthétiques générées à partir de l'IRM UTE pourraient potentiellement répondre aux normes requises pour une utilisation clinique. Les images synthétiques correspondaient étroitement aux images CT réelles en termes de précision, notamment dans l'évaluation des altérations structurelles liées à la mucoviscidose. Cette évaluation a également révélé que les images CT synthétiques permettaient une meilleure analyse clinique des lésions, telles que les bronchiectasies, par rapport à l'IRM UTE, renforçant la viabilité de cette approche comme alternative sans radiation pour l'imagerie pulmonaire.
Cependant, l'étude a également identifié certaines limitations. Bien que les images synthétiques soient prometteuses, le potentiel des GAN à introduire des caractéristiques hallucinées demeure une préoccupation, particulièrement dans des contextes cliniques. De plus, étant donné que les structures d'intérêt telles que les bronches et les vaisseaux sanguins sont intrinsèquement tridimensionnelles, l'approche actuelle en 2D peut ne pas capturer pleinement les relations spatiales et le contexte nécessaires pour une synthèse précise. Cela souligne la nécessité d'introduire des informations 3D pour mieux exploiter les données de voisinage et améliorer la représentation des structures anatomiques complexes.
La dépendance aux métriques de traitement d'images standard, qui peuvent ne pas refléter pleinement la précision anatomique requise en imagerie médicale, souligne également le besoin de méthodes d'évaluation plus pertinentes sur le plan clinique. Les métriques basées sur l'intensité ou la similarité globale ne capturent pas toujours les détails anatomiques critiques pour le diagnostic. Par conséquent, de futures recherches devraient se concentrer sur le développement de métriques automatisées capables d'évaluer de manière fiable la précision des reconstructions anatomiques, réduisant ainsi la dépendance aux évaluations manuelles par des radiologues experts.
En somme, ce chapitre démontre la faisabilité de la synthèse d'images CT à partir de l'IRM UTE du poumon en utilisant des GAN 2D, tout en mettant en évidence les défis et les domaines nécessitant une attention future. L'optimisation du prétraitement des données, le choix judicieux des modèles génératifs et l'évaluation clinique rigoureuse sont essentiels pour progresser vers une application clinique fiable de cette technologie. Les prochaines étapes impliqueront l'intégration d'informations 3D et le développement de métriques d'évaluation spécifiques à la tâche pour améliorer encore la qualité et la pertinence clinique des images synthétiques.

Dans le Chapitre 3, nous avons approfondi l'optimisation des approches de synthèse

d'images de l'IRM vers le scanner CT, en mettant l'accent sur l'amélioration de la reconstruction des structures bronchiques. Cette focalisation découle de l'importance clinique des bronches dans le diagnostic et le suivi des maladies pulmonaires, ainsi que des défis techniques liés à la capture de ces structures fines en imagerie synthétique.
Tout d'abord, nous avons réévalué les métriques de traitement d'images traditionnelles, en soulignant leurs limites dans le contexte de la synthèse d'images médicales. Les métriques basées sur l'intensité, telles que l'erreur quadratique moyenne (MSE) et le pic du rapport signal sur bruit (PSNR), ne reflètent pas la qualité de la reconstruction des structures anatomiques fines comme les voies aériennes. Pour pallier cette insuffisance, nous avons introduit des métriques spécifiques à une tâche donnée, notamment le coefficient de Dice et la Distance de Surface Normalisée (NSD), adaptées pour évaluer la qualité des reconstructions bronchiques. En utilisant la pipeline de segmentation NaviAirway, nous avons pu quantifier de manière plus pertinente la performance de nos modèles, en nous concentrant sur la précision anatomique plutôt que sur la simple similarité d'intensité.
Ensuite, nous avons optimisé l'utilisation de la fonction de coût perceptuelle pour la synthèse 2D de l'IRM vers le CT. En menant une investigation systématique sur l'impact de différents paramètres, y compris les techniques de prétraitement, l'architecture du réseau et la sélection des couches, nous avons démontré que l'utilisation de ResNet50 pré-entraîné sur ImageNet, combiné à une fonction de coût perceptuelle contrastive, améliorait significativement la qualité des images synthétisées. Cette approche a permis une meilleure préservation des structures bronchiques par rapport aux fonctions de coût traditionnelles. Cependant, nous avons également identifié une limitation majeure : la fonction de coût perceptuelle actuelle est limitée aux calculs en 2D et peut ne pas capturer pleinement la complexité 3D inhérente à l'imagerie médicale.
Pour surmonter cette limitation, nous avons introduit et comparé plusieurs approches 2.5D visant à incorporer des informations 3D dans les réseaux conçus pour le traitement 2D. La stratégie de fusion multi-axiale et l'approche des tranches empilées ont toutes deux démontré une amélioration de la qualité des images et de la reconstruction des voies aériennes par rapport à la référence 2D. Cependant, la méthode la plus efficace s'est avérée être notre approche hybride proposée. En combinant la fusion multi-axiale avec les tranches empilées, nous avons atteint les meilleures performances dans la reconstruction des structures anatomiques fines. Cette méthode hybride a permis l'intégration d'informations 3D sans le lourd fardeau computationnel associé aux modèles 3D.
En conclusion, ce chapitre a démontré que l'intégration de métriques spécifiques à la reconstruction anatomique, l'optimisation de la fonction de coût perceptuelle et l'adoption de stratégies 2.5D peuvent considérablement améliorer la synthèse d'images IRM vers CT, notamment pour la reconstruction des structures bronchiques. Ces avancées contribuent à combler l'écart entre la qualité des images synthétiques et les exigences cliniques, rapprochant ainsi cette technologie d'une application pratique en milieu clinique.

Dans le Chapitre 4, nous avons élargi notre focus au-delà de la synthèse IRM pulmonaire vers CT pour développer une méthode plus généralisable applicable à diverses tâches de synthèse d'images médicales. En utilisant le cadre nnU-Net, nous avons créé une solution évolutive capable de relever différents défis de synthèse à travers plusieurs régions anatomiques. Cette approche vise à surmonter les limitations des méthodes précédentes et à établir une base solide pour la synthèse d'images médicales de haute qualité.
Pour valider notre méthode, nous l'avons appliquée à des jeux de données publics pour des tâches telles que la conversion IRM vers CT du pelvis et l'inpainting d'IRM cérébrales.

En adaptant l'architecture nnU-Net pour la synthèse d'images médicales en 3D, nous avons établi une référence reproductible et cohérente, surpassant les approches basées sur les GAN, qui nécessitent souvent un entraînement antagoniste complexe. Grâce à sa configuration automatique et à son optimisation spécifique aux caractéristiques du jeu de données utilisé, nnU-Net s'est avéré plus flexible, plus léger et plus efficace que les modèles de diffusion, ce qui en fait une solution pratique pour la synthèse d'images médicales dans diverses tâches et régions anatomiques.
Cependant, nous avons identifié des limitations lors de l'utilisation de fonctions de coût standards, telles que L1 et la fonction de coût perceptuelle 3D, qui peinent à capturer les détails anatomiques fins. Ces fonctions de coût traditionnelles se concentrent principalement sur les différences globales d'intensité ou sur la similarité perceptuelle générale, sans nécessairement accorder une attention particulière aux structures anatomiques critiques. Cette limitation a motivé le développement de notre fonction de coût priorisant les caractéristiques anatomiques, appelée AFP Loss (Anatomical Feature-Prioritized Loss).
L'AFP Loss représente une avancée significative en intégrant des connaissances spécifiques au domaine dans le processus de synthèse. En exploitant les représentations issues de réseaux de segmentation pré-entraînés, l'AFP Loss permet une reconstruction ciblée des caractéristiques anatomiques essentielles telles que les voies aériennes, les organes, les muscles et les os. Cette priorité accordée aux structures critiques améliore la capacité du modèle à reconstruire avec précision les détails anatomiques importants, ce qui est essentiel pour l'applicabilité clinique des images synthétiques.
Nous avons démontré que l'intégration de l'AFP Loss améliore non seulement les tâches de synthèse intra-pulmonaires, mais montre également son efficacité dans des zones extra-pulmonaires. En orientant le modèle pour qu'il accorde une importance accrue aux structures d'intérêt, nous avons réussi à améliorer la qualité globale des images synthétiques, tout en maintenant une cohérence avec les données réelles.
En conclusion, ce chapitre présente une approche complète de la synthèse d'images médicales qui combine la flexibilité de nnU-Net avec la précision anatomique de l'AFP Loss. Notre travail améliore fondamentalement la précision et la reproductibilité des images médicales synthétisées, offrant une solution évolutive pour diverses tâches d'imagerie médicale. La validation de nos méthodes sur des jeux de données publics souligne leur généralisabilité et leur potentiel clinique.

En conclusion, cette thèse a démontré le potentiel de l'apprentissage profond pour améliorer l'imagerie pulmonaire, en combinant la sécurité de l'IRM avec la précision du scanner TDM. Les avancées réalisées ouvrent la voie à des soins plus sûrs et efficaces, et posent les bases pour des recherches futures visant à intégrer davantage de connaissances spécifiques au domaine dans les modèles d'apprentissage profond, améliorant ainsi leur applicabilité clinique.
Les perspectives futures de ce travail sont multiples. Premièrement, le raffinement de l'AFP Loss par l'intégration de stratégies de pondération dynamique pourrait améliorer l'adaptabilité du modèle à divers jeux de données et tâches, réduisant le besoin d'ajustements manuels. De plus, l'incorporation de connaissances physiques, comme les données du domaine fréquentiel en IRM, pourrait enrichir le modèle en lui fournissant une compréhension plus profonde de la formation des images, améliorant ainsi la fidélité des images synthétiques.

Publications

Journal Papers

[1] **A. Longuefosse**, J. Raoult, I. Benlala, et al. "Generating high-resolution synthetic CT from lung MRI with ultrashort echo times: initial evaluation in cystic fibrosis". *Radiology*, 308 (2023).

Submitted Journal Papers

[2] **A. Longuefosse**, B. Denis de Senneville, G. Dournes, et al. "Anatomical feature-prioritized loss for enhanced MR to CT translation". Under revision in *Computers in Biology and Medicine*.

International Conference Proceedings

[3] **A. Longuefosse**, G. Dournes, I. Benlala, et al. "Lung CT Synthesis Using GANs with Conditional Normalization on Registered Ultrashort Echo-Time MRI". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2023)*.

[4] **A. Longuefosse**, B. Denis de Senneville, G. Dournes, et al. "MR to CT synthesis using GANs: a practical guide applied to thoracic imaging". In *International Conference on Information Visualization Theory and Applications, IVAPP (2023)*.

[5] **A. Longuefosse**, B. Denis de Senneville, G. Dournes, et al. "Incorporating 3D information in 2.5D networks: strategies for MR to CT synthesis". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2024)*.

[6] **A. Longuefosse**, B. Denis de Senneville, G. Dournes, et al. "On the use of perceptual loss for fine structure generation: Illustration on lung MR to CT synthesis". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2024)*.

[7] **A. Longuefosse**, E. Le Bot, B. Denis de Senneville, et al. "Adapted nnU-Net: A Robust Baseline for Cross-Modality Synthesis and Medical Image Inpainting". In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), 2024, Held in Conjunction with MICCAI 2024*.

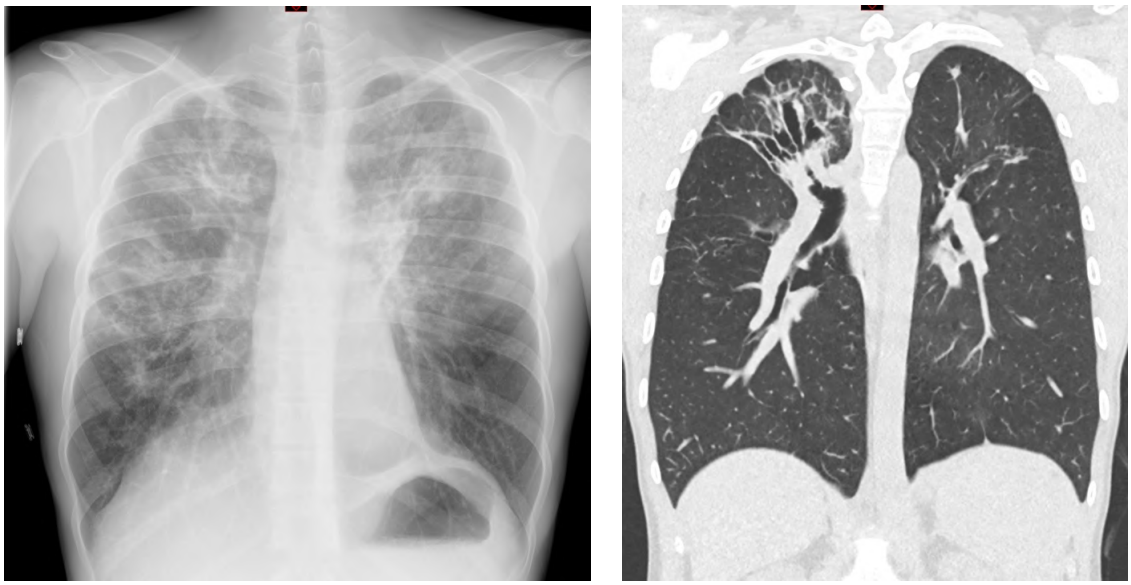
Introduction

Clinical Context of Lung Imaging

Lung imaging plays a pivotal role in the diagnosis, treatment planning, and monitoring of various pulmonary diseases, such as chronic obstructive pulmonary disease (COPD), lung cancer, cystic fibrosis (CF), and interstitial lung diseases like idiopathic pulmonary fibrosis (IPF). Accurate and detailed imaging is essential for evaluating disease progression, guiding interventions, and improving patient outcomes.
Several imaging modalities are utilized in clinical practice for lung imaging, each with its own advantages and limitations.

Chest Radiography Chest radiography is a widely accessible, fast, and cost-effective imaging technique commonly used as the initial tool for assessing lung conditions. It involves low levels of radiation and can detect a range of pulmonary abnormalities, such as pneumothorax, pneumonia, and pleural effusions. Despite its advantages, chest radiographs have limitations in sensitivity and specificity, particularly for early-stage or subtle lung diseases, due to their two-dimensional nature and limited contrast resolution. Additionally, the interpretation of chest radiographs can be subject to variability among different observers, which may impact diagnostic consistency.

Computed Tomography Computed tomography (CT) is considered the gold standard for lung imaging due to its high spatial resolution and excellent contrast differentiation between pulmonary structures. CT scans provide detailed 3D images, allowing for precise evaluation of lung anatomy, detection of small nodules, assessment of airway abnormalities, and quantification of disease extent. In conditions like cystic fibrosis and interstitial lung diseases, CT is invaluable for detecting and quantifying bronchiectasis, fibrotic changes, and other subtle anomalies. Scoring systems such as the Bhalla or Brody scores offer objective and reproducible assessments of disease severity based on CT findings. Another significant advantage of CT is its rapid acquisition time, typically ranging from 5 to 20 seconds. Figure 1 illustrates a comparison between chest radiography and CT, demonstrating the superior anatomical detail provided by the CT scan.
Despite its diagnostic advantages, CT imaging involves exposure to ionizing radiation, which raises concerns, particularly for pediatric patients and individuals requiring frequent imaging, such as those with chronic lung diseases. Repeated radiation exposure increases the cumulative risk of radiation-induced malignancies. The U.S. Food and Drug Administration (FDA) recognizes that exposure to a dose of 10 msv from a CT scan may



X-ray CT scan

Figure 1: Comparison of radiographic imaging modalities in a patient with cystic fibrosis. The CT scan provides a more detailed, 3D cross-sectional view, particularly highlighting bronchiectasis and ventilation abnormalities.

result in a 1 in 2000 chance of developing a fatal cancer. Efforts to mitigate this risk have led to the development of low-dose and ultra-low-dose CT protocols. While these techniques reduce radiation exposure significantly, they may involve trade-offs in image quality, particularly in the visualization of fine anatomical details.
Another limitation of CT imaging is the variability in lung volume measurements due to the requirement of breath-holding during acquisition. Differences in inspiratory effort between scans can affect the reproducibility of volumetric assessments, which is critical in monitoring disease progression.

Magnetic Resonance Imaging Magnetic resonance imaging (MRI), despite its excellent soft tissue contrast and functional imaging capabilities, has historically been underutilized for lung imaging due to the technical challenges presented by the lung's unique structure. Classical MRI sequences, such as T1 and T2-weighted imaging, face significant limitations in the lungs due to the low proton density of air-filled spaces, resulting in weak signal intensity. Additionally, the numerous air-tissue interfaces cause magnetic field inhomogeneities, leading to susceptibility artifacts and rapid signal decay. The longer acquisition times typically required for MRI, ranging from 5 to 20 minutes, often necessitate free-breathing during the scan, which introduces motion artifacts caused by respiratory and cardiac motion. While T2-weighted imaging can be used to assess bronchial lesions, edema, masses, and other forms of inflammation, its application in lung imaging remains limited. These technical challenges, combined with the susceptibility to motion artifacts, have restricted the use of conventional MRI in pulmonary diagnostics, particularly when compared to other modalities like CT, which offers superior spatial resolution, faster acquisition, and lower cost, making it the preferred option in clinical practice.

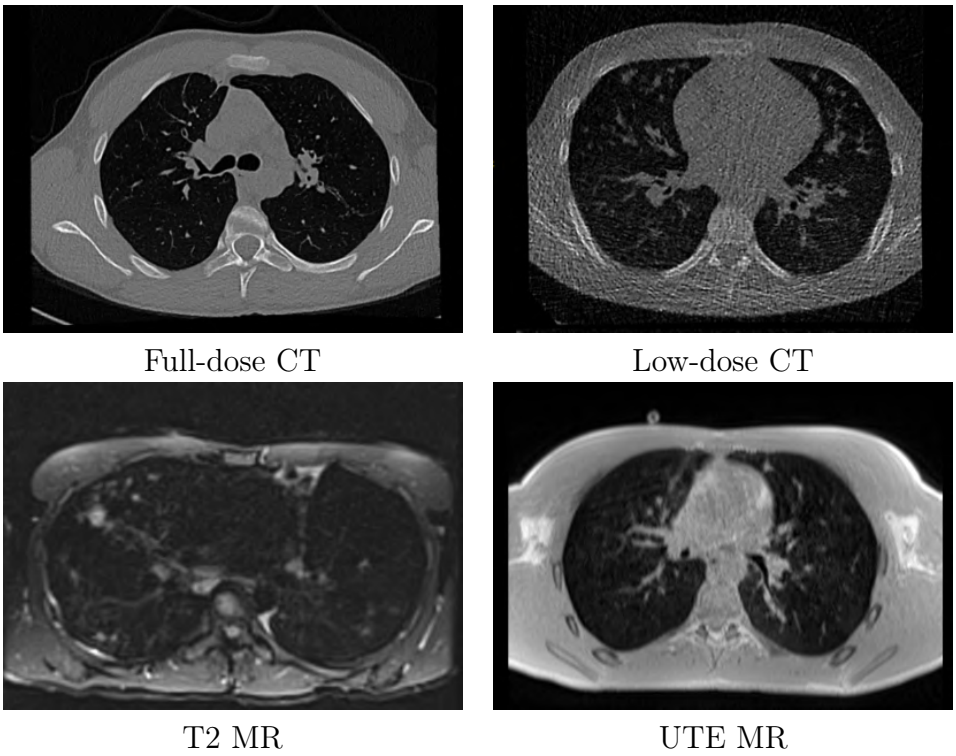


Figure 2: Visual comparison of axial slices across four modalities: Full-dose CT with high-resolution and signal-to-noise ratio (SNR), low-dose CT with increased noise, T2-weighted MRI with low spatial resolution, and UTE MRI with increased structural details compared to traditional MRI sequences.

Current Solutions in Pulmonary Imaging

To balance the need for detailed lung imaging with the imperative to minimize radiation exposure, several strategies have been developed in clinical practice. One primary solution is the use of low-dose and ultra-low-dose computed tomography (CT) scans. Technological advancements have enabled significant reductions in radiation dose without proportionally compromising image quality. Techniques such as automatic exposure control and iterative reconstruction algorithms allow for dose reductions while maintaining sufficient diagnostic information. Low-dose CT scans deliver significantly less radiation compared to conventional CT scans, making them more suitable for populations requiring frequent imaging.

However, reducing the radiation dose is not without drawbacks. Lowering the dose can lead to increased image noise and artifacts, potentially obscuring fine anatomical details crucial for accurate diagnosis. Photon starvation and beam hardening artifacts may become more pronounced, affecting the visualization of subtle lesions or small airways. As a result, while low-dose CT scans mitigate the risks associated with radiation exposure, they may also compromise the quality of the images obtained, which can hinder the detection and characterization of pulmonary pathologies. In parallel, magnetic resonance imaging (MRI) has emerged as a promising radiation-free alternative for lung imaging. Recent developments in MRI technology, such as the implementation of ultrashort echo time (UTE) sequences, have significantly improved the modality's capacity to visualize pulmonary structures. Sequences like the three-dimensional Volume Interpolated Breath-hold Examination (3D VIBE) UTE enable the acquisition of high-resolution images with

millimetric precision, enhancing the visualization of lung parenchyma and airways.
These advanced MRI sequences address some of the traditional limitations of lung MRI by capturing signals from tissues with very short T2* relaxation times, which is essential in the lung due to its low proton density and susceptibility-induced signal decay. The 3D VIBE UTE sequence, for example, allows for detailed morphological assessment of the lungs, providing valuable information for the evaluation of various pulmonary diseases.
Despite these advancements, MRI still faces challenges in lung imaging. The spatial resolution and signal-to-noise ratio, although improved, generally remain inferior to those of CT scans. Additionally, MRI is more susceptible to artifacts arising from respiratory and cardiac motion, which can blur images and obscure fine details. The longer acquisition times required for MRI compared to CT can also be a limitation, potentially causing discomfort for patients and increasing the likelihood of motion artifacts.
Therefore, while both low-dose CT and advanced MRI techniques like UTE sequences offer alternatives to conventional imaging, each has inherent limitations that can affect image quality and diagnostic accuracy. Clinicians must carefully consider these factors when choosing the most appropriate imaging modality for their patients.

Motivation for MRI to CT Translation

The idea of combining the advantages of MRI and CT imaging modalities has led to the exploration of transforming MRI data into CT-like images through advanced computational methods. The motivation behind this approach is to leverage the radiation-free nature of MRI while attaining the high-resolution anatomical detail characteristic of CT scans.
MRI to CT translation holds significant potential benefits. It could enable the generation of high-quality, CT-equivalent images without exposing patients to ionizing radiation, thus enhancing patient safety. This approach could also improve patient comfort by reducing the need for breath-holding during scans, as MRI can be performed using free-breathing techniques.
Clinically, accurate MRI to CT translation would expand the utility of MRI in pulmonary diagnostics, allowing for detailed visualization of pulmonary structures, including small airways and vessels, which are critical for the diagnosis and management of various lung diseases. This could facilitate better monitoring of disease progression, assessment of treatment efficacy, and early detection of complications.
However, achieving reliable MRI to CT translation is technically challenging. The fundamental differences in the physical principles underlying MRI and CT imaging result in different contrast mechanisms and image characteristics. MRI captures proton density and tissue relaxation properties, while CT measures X-ray attenuation coefficients. Accurately mapping MRI signal intensities to corresponding CT attenuation values requires sophisticated algorithms capable of capturing complex relationships between the modalities.
Moreover, preserving fine anatomical details during the translation process is essential for clinical applicability. Any loss of detail or introduction of artifacts could compromise diagnostic accuracy. The variability in MR image quality due to factors such as patient movement, differences in scanning protocols, and magnetic field inhomogeneities adds another layer of complexity to the translation task.
Recent advances in deep learning and image synthesis techniques offer promising solutions to these challenges. Convolutional neural networks and generative adversarial networks

have shown potential in learning complex mappings between MRI and CT images, enabling the generation of synthetic CT images that closely resemble real CT scans. These approaches aim to maintain anatomical fidelity while leveraging the advantages of MRI. In this context, a collaborative effort between the Centre Hospitalier de Bordeaux and the Laboratoire Bordelais de Recherche en Informatique (LaBRI) has been established to address these challenges using deep learning techniques. This interdisciplinary collaboration combines clinical expertise in thoracic imaging with advanced computational methodologies in artificial intelligence and machine learning. The partnership aims to develop robust, efficient, and accurate models for MR to CT translation specifically tailored to thoracic imaging.

Contributions and Thesis Structure

This thesis addresses the challenges of transforming magnetic resonance imaging (MRI) data into computed tomography (CT)-like images for thoracic imaging. The primary motivation is to leverage the radiation-free nature of MRI while obtaining the high-resolution anatomical details characteristic of CT scans, thereby enhancing patient safety and diagnostic capabilities in pulmonary imaging.
The contributions of this thesis are multifaceted, encompassing the development and optimization of deep learning models for MR to CT translation, the introduction of novel techniques to improve anatomical reconstruction during synthesis, and the comprehensive evaluation of these methods in clinical contexts. The work is structured into four main chapters, each building upon the previous to advance the state-of-the-art in medical image synthesis for lung imaging.

Chapter 1 lays the groundwork by exploring the fundamental concepts of image synthesis using traditional and deep learning techniques, with a particular emphasis on architectures used in image translation, and their application in medical imaging. Furthermore, it discusses the unique challenges present in lung imaging, including the variability in lung anatomy, the presence of fine structures like bronchi and vessels, and the impact of respiratory motion on image acquisition.

Chapter 2 presents our first study towards the feasibility of using 2D GANs for MR to CT translation in the lungs. This chapter explores the advantages of using spatially adaptive normalization techniques to preserve spatial structures during the synthesis process, which is crucial for accurate lung imaging. It also discusses the impact of optimizing preprocessing techniques, such as resampling and image registration, to improve image quality and synthesis accuracy. The chapter concludes with an assessment of the clinical viability of these methods, supported by a study with radiologists to validate the synthesized images.

Chapter 3 focuses on improving MR to CT synthesis by refining both the evaluation pipeline and the translation process. We emphasize preserving fine anatomical structures like vessels and bronchi, introducing task-specific metrics using segmentation networks to assess their reconstruction. The synthesis is optimized through the use of perceptual loss functions and 2.5D approaches. Perceptual loss improves anatomical accuracy by comparing high-level image features. The 2.5D approach integrates 3D context by processing

slices from multiple planes, offering a balance between the efficiency of 2D models and the spatial coherence of 3D models, providing a lightweight yet effective alternative to full 3D models.

Chapter 4 explores advanced 3D synthesis models and innovative techniques to address the specific challenges of MR to CT translation for lung imaging. The adaptation of the nnU-Net framework, originally designed for segmentation tasks, is presented as a robust and scalable baseline for cross-modality synthesis. We then introduce the Anatomical Feature-Prioritized (AFP) loss, which prioritizes the reconstruction of critical anatomical structures such as airways and vessels over global intensity matching. The chapter concludes with an evaluation of the proposed methods on public datasets, demonstrating the efficacy of the adapted nnU-Net framework and the AFP loss in generating high-quality synthetic images while preserving anatomical details.

Chapter 1

Fundamentals of Medical Image Synthesis

In the previous section, we outlined the clinical context and the inherent challenges in lung imaging, and our approach to address this by synthesizing CT scans from lung UTE MR. This first chapter lays the foundational concepts around medical image synthesis, essential to understanding the methods and challenges explored in the subsequent chapters of this thesis.
The chapter opens with an overview of the evolution of image synthesis, starting with traditional approaches and progressing to recent deep learning-based methods. This section highlights the key techniques and methodologies that have shaped the field over time, along with various applications of image synthesis in medical imaging, focusing on the generation of synthetic data for enhancing clinical workflows. Following this, we focus on image-to-image translation, reviewing state-of-the-art architectures and their applications, particularly in cross-modality medical translation tasks, such as MR to CT synthesis. We then present the key evaluation metrics used to assess the quality of synthesized images, discussing both general image quality metrics and GAN-specific measures. We also examine the challenges that persist in evaluating medical image synthesis, emphasizing the clinical relevance of these assessments.
Lastly, we address the limitations of image synthesis in thoracic imaging, specifically related to the challenges posed by current datasets, the quality of lung UTE MR data, and the complexity of thoracic structures such as airways and vessels. These limitations form the basis for the methods and solutions explored in the subsequent chapters.

1.1 Introduction to Image Synthesis

1.1.1 Traditional Approaches

Prior to the advent of deep learning, various traditional approaches were employed for synthesizing medical images. These methods laid the groundwork for modern techniques and can be broadly categorized into three major families: acquisition-based approaches, transform-based approaches, and exemplar-based approaches, as outlined in [1]. Each of these approaches offered a different perspective on how to generate or modify images based on available data, computational models, and image statistics.

Acquisition-based

Acquisition-based methods for medical image synthesis focus on capturing intrinsic MRI tissue parameters such as T1, T2, and proton density, allowing for the synthesis of any desired MRI contrast. These approaches acquire multiple images with varying parameters, such as different repetition times and echo times, to sample the relaxation curves of tissues. The acquired data is then fit to MRI signal behavior models to extract tissue-specific parameters, which are subsequently used to generate synthetic contrasts.
A key early contribution to this field was made by Bobman et al. in 1985 [2], who demonstrated the potential of synthesizing MRI contrasts from parametric maps of tissue properties. This work laid the foundation for using physical MRI parameters to generate synthetic images, opening the door to more sophisticated acquisition-based methods.
In 2004, another significant advancement was introduced with the development of a sequence-independent approach for segmentation and registration of MR images [3]. This method embedded the physics of the imaging process into the segmentation or registration algorithm, making it invariant to acquisition parameters. By explicitly modeling how acquisition parameters affected tissue contrast, the technique allowed for the generation of probabilistic models of intrinsic tissue parameters, such as T1, proton density, and T2*. This approach offered the additional benefit of detecting physiologically relevant changes in tissue parameters that are often difficult to detect in traditional contrast-weighted MR images. Figure 1.1 illustrates an example of an acquisition-based pipeline for brain image synthesis.

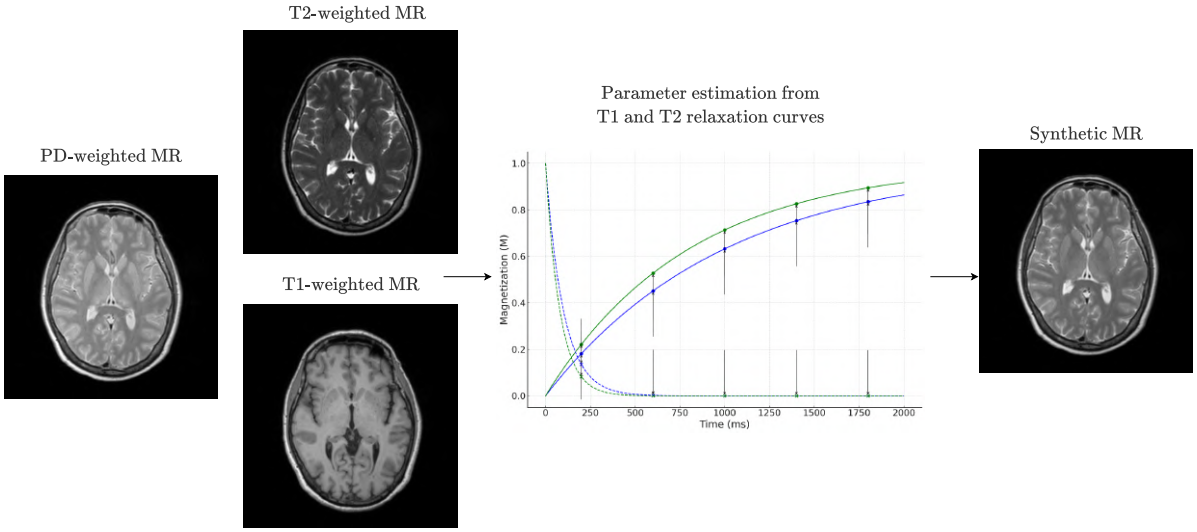


Figure 1.1: Example of MR synthesis with quantified tissue parameters using brain images from the public IXI dataset [4]. Proton density (PD), T1-weighted, and T2-weighted MR images are acquired to estimate T1, T2, and PD parameters. These quantified parameters are used to generate synthetic MR images, allowing multiple contrasts from a single acquisition.

Building on these early developments, Warntjes et al. (2008) introduced synthetic MRI techniques optimized for rapid quantification of multiple tissue parameters [5]. By reducing the number of required acquisitions, they made the process more efficient while still capturing the essential information needed for accurate tissue characterization.
Despite these advancements, acquisition-based approaches are constrained by long acqui-

sition times, susceptibility to motion artifacts, and the need for specialized sequences and hardware. The accuracy of the synthesized images remains highly dependent on precise parameter estimation, and errors in this process can degrade the quality of the images, limiting the use of these techniques in routine clinical practice.

Transform-based

Transform-based approaches aim to model the relationship between different imaging modalities or contrasts through mathematical transformations or statistical models. These methods operate by learning a mapping between input and target modalities, often using a single input image to estimate the target image. Unlike acquisition-based methods, transform-based approaches focus on developing a generalizable model that can map one contrast to another, capturing the underlying relationship between different imaging modalities.
One common type of transform-based approach uses Gaussian Mixture Models (GMMs) to model the joint intensity distribution between different modalities [6]. By fitting the intensity values of a given image to a GMM, the algorithm can generate synthetic images by transforming the input modality based on the learned relationships between intensity values. Another frequently used technique involves polynomial transforms, where parametric function are applied to map the intensities of one modality to another. These parametric models are highly flexible and can be tailored to capture specific relationships between modalities. Guimond et al. (2001) applied polynomial transformations alongside the Demons algorithm for multi-modal brain warping, demonstrating the utility of these models for aligning different imaging modalities [7]. Kroon et al. (2009) expanded on this by incorporating a non-parametric joint histogram model, which better captured intensity relationships during MRI modality transformation [8].
While these approaches are generalizable, they also face challenges. Complex or highly non-linear relationships between imaging modalities can be difficult to model accurately using simple transformations or statistical models. As a result, transform-based methods may struggle to achieve the precision necessary for highly detailed tasks, such as the synthesis of complex anatomical structures. Furthermore, the accuracy of the synthesized images is often limited by the chosen model's complexity, and careful parameter tuning is typically required to optimize the results.

Exemplar-based

Exemplar-based approaches are grounded in the use of a pre-acquired database of images, typically from different subjects, where both source and target modalities are available. The underlying principle is to use the source modality of a test subject, for which a target needs to be synthesized, and match it to the corresponding information in the database to generate the target image. These methods often leverage either atlas-based or patch-based techniques, which naturally incorporate spatial context and are highly generalizable across datasets.

Atlas-based methods Atlas-based methods involve the use of a database of paired images, often in the form of CT and MR scans, where the anatomical structures are pre-aligned. The atlas, once registered to a new subject's MR image, provides the corresponding CT data through rigid or non-rigid transformation, allowing the synthesis of a CT image for the new subject. This approach has been particularly valuable in scenarios

where detailed anatomical alignment is crucial, such as in brain imaging or radiotherapy planning. For instance, Roy et al. (2013) employed atlas-based techniques for MR image contrast synthesis, enhancing the ability to generate missing MR contrasts by mapping from existing ones [9]. Figure 1.2 provides an illustration of an atlas-based pipeline for pelvic MR to CT synthesis.

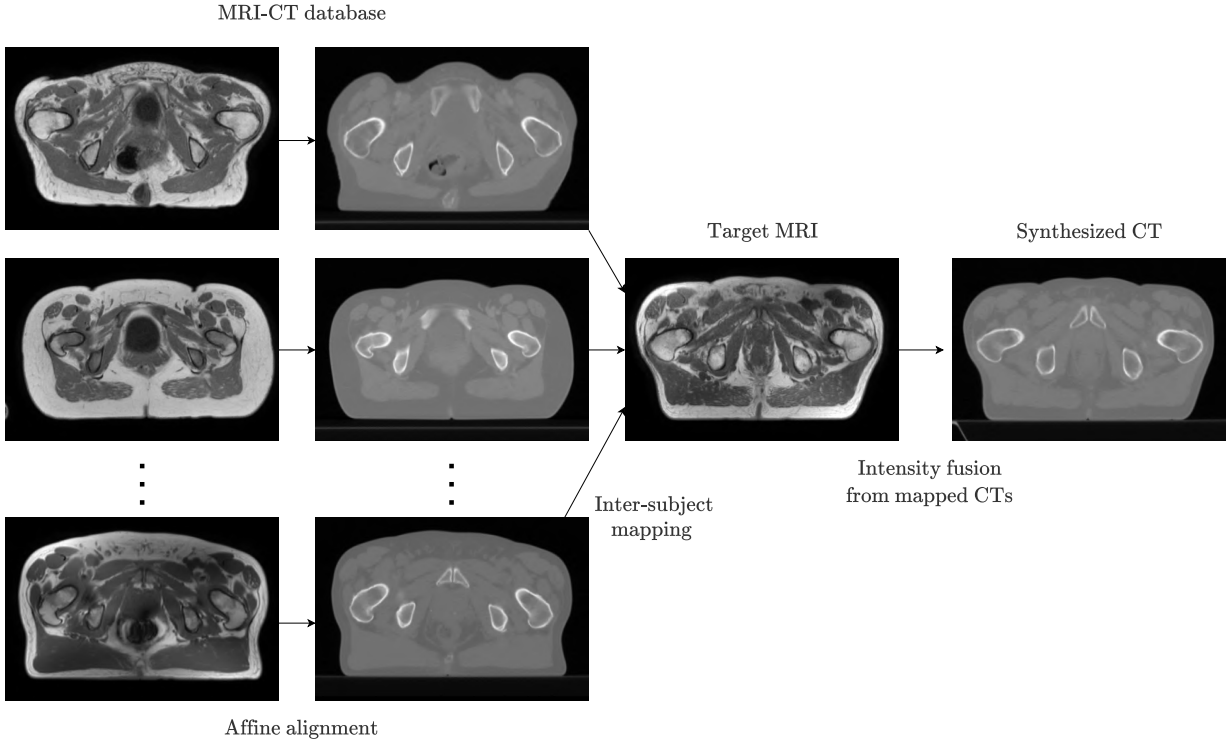


Figure 1.2: Example of CT synthesis in the pelvis using an atlas-based method. All MR images in the database are registered to the target MR image. The CTs are then mapped using the same transformation to the target MR image. The target CT is synthesized using an intensity fusion from mapped CTs.

Burgos et al. (2014, 2015) also made significant contributions to atlas-based synthesis, applying it to CT synthesis for PET/MR attenuation correction and radiotherapy planning [10, 11]. In their work, multi-atlas registration was used to transform and fuse CT images from an atlas onto the target MR image, allowing the generation of synthetic CT scans that accurately represent the subject's anatomy. This method proved highly effective for head and neck imaging, where accurate registration of complex anatomical regions is critical for clinical applications.
Despite their robustness, atlas-based methods are heavily dependent on the quality and diversity of the atlas. Poor matches between the subject and atlas can result in inaccuracies, and the computational complexity of multiple registrations can be a limitation.

Patch-based methods Patch-based methods take a more localized approach, synthesizing target images by finding similar patches in a database of example image pairs. Instead of relying on global registration like atlas-based methods, patch-based approaches identify small regions (patches) in the input image, search for matching patches in the database, and then use the corresponding patches from the target modality for synthesis. Patch-based methods are particularly effective in capturing local image characteristics

and handling fine details, especially when the relationship between modalities is highly complex or non-linear. For instance, Ye et al. used a patch-based approach for modality propagation, where subject-specific scans were synthesized with data-driven regularization, allowing for coherent image generation [12]. Similarly, Rousseau et al. (2011) employed a supervised patch-based approach for brain labeling, showing the method's effectiveness in segmentation tasks [13].
One of the key advantages of patch-based methods is their ability to work without the need for perfect global registration, making them more flexible when handling local variations in image intensity and structure. However, they can be computationally intensive, particularly when searching for matching patches in large databases, and inconsistencies at patch boundaries can sometimes arise, resulting in visible artifacts in the synthesized images.

Both atlas-based and patch-based methods leverage a database of pre-existing image pairs, providing flexibility in handling complex anatomical variations and cross-modality relationships. However, these approaches are highly dependent on the availability and quality of the exemplar databases. Atlas-based methods often struggle with cases where anatomical differences between the subject and the atlas are significant, leading to inaccuracies. Patch-based methods, while capturing local image characteristics, can suffer from artifacts at patch boundaries and require extensive computational resources for patch matching, particularly in large datasets.
These limitations, combined with the increasing complexity of modern medical imaging tasks, highlight the need for more advanced techniques such as machine learning and deep learning, which capture intricate patterns directly from the data, and provide improved generalization across diverse imaging scenarios.

1.1.2 Machine Learning Approaches

Early machine learning approaches for medical image synthesis primarily focused on methods such as random forests, support vector machines (SVM), and gradient boosting, aiming to learn complex non-linear mappings between different imaging modalities or contrasts without relying on deep neural networks. These techniques demonstrated the potential of data-driven models to handle the intricacies of medical imaging and produce synthetic images, albeit with certain limitations.
Random forests, in particular, gained traction due to their ability to manage high-dimensional data and capture complex relationships between input and output spaces. For example, Jog et al. (2017) proposed a random forest regression approach to synthesize T2-weighted and FLAIR images from T1-weighted MRI, showing reasonable accuracy in intensity prediction for different modalities [14]. Similarly, Huynh et al. (2016) utilized structured random forests for estimating CT from MRI for PET/MR attenuation correction, which allowed for improved diagnostic workflows without the need for additional CT scans [15]. In the same vein, support vector regression was explored by Vemulapalli et al. (2015) for cross-modality synthesis, demonstrating the application of kernel methods to learn mappings between modalities [16].
These methods, while effective in capturing non-linear relationships, had notable limitations. Their reliance on handcrafted features and predefined intensity mappings often restricted their ability to generalize across more complex anatomical structures and larger datasets. Additionally, their performance was dependent on the quality and amount of

training data, and they often required significant manual tuning to achieve optimal results. While these machine learning techniques laid the groundwork for medical image synthesis, their limitations highlighted the need for more powerful, automated approaches capable of learning deeper representations of the data.

The Rise of Deep Learning in Medical Imaging

Deep learning, a subset of machine learning, has brought significant advancements in various domains, particularly in image analysis. Unlike traditional machine learning approaches that rely on handcrafted features, deep learning allows models to automatically learn hierarchical representations of data through multiple layers of abstraction [17]. These models, especially neural networks, have demonstrated their ability to handle complex, high-dimensional medical imaging data, uncovering patterns that were previously difficult to model using traditional methods [18]. With advances in hardware and computational power, deep learning has become the go-to approach for medical image analysis, laying the foundation for further innovations in image synthesis. Deep learning, particularly through convolutional neural networks (CNNs), has transformed image analysis by enabling models to automatically learn spatial hierarchies from data [19]. CNNs, through their layered architecture presented in Figure 1.3, can capture local image features and progressively build representations of increasing complexity.

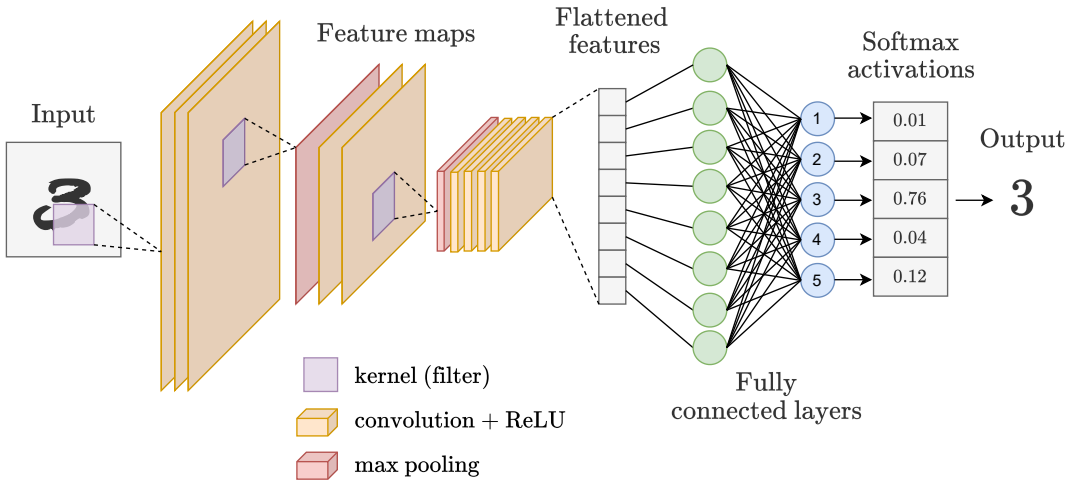


Figure 1.3: Example of CNN architecture for handwritten digit classification.

In the context of medical imaging, these models initially gained traction for tasks such as image classification, disease diagnosis, semantic segmentation, and have also found applications in image synthesis. For example, the work by Li et al. (2014) demonstrated the utility of CNNs in improving diagnostic workflows by completing missing imaging data [20].

From CNNs to U-Net

A key advancement in CNN-based architectures for medical applications was the development of U-Net by Ronneberger et al. (2015), which introduced the concept of *skip connections* [21]. These connections link the contracting (encoder) and expanding (decoder) paths of the network, allowing high-resolution features from earlier layers to be

directly incorporated into later layers. This mitigates the loss of spatial information during downsampling, which is crucial for pixel-accurate tasks such as segmentation. U-Net's ability to localize and distinguish fine-grained structures made it particularly suited for medical image tasks, where the accurate delineation of anatomical boundaries is essential. Figure 1.4 presents an overview of a standard U-Net architecture.

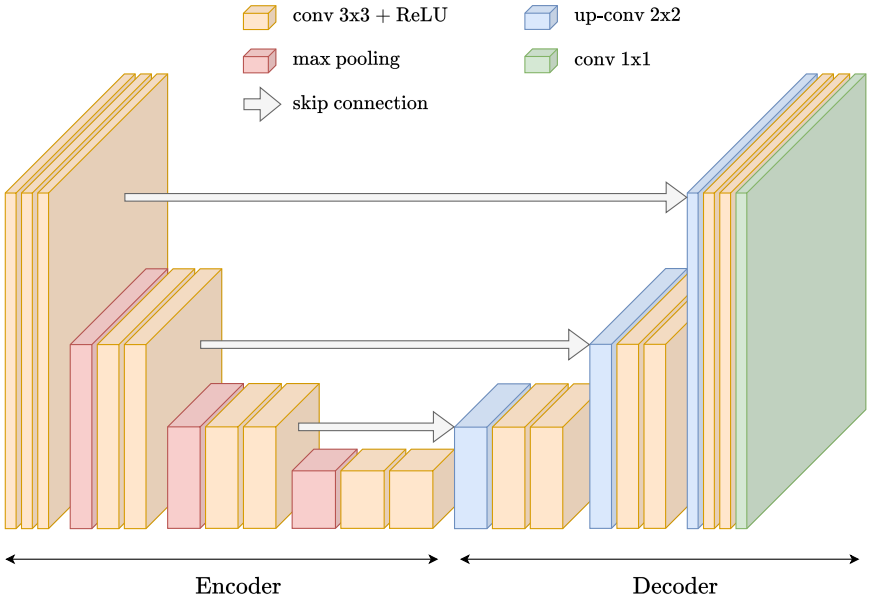


Figure 1.4: Example of standard U-Net architecture.

Initially developed for semantic segmentation, U-Net quickly found applications in medical image synthesis due to its powerful ability to handle the complexities of medical data. A notable example is its use in whole brain segmentation and labeling from CT using synthetic MR images [22], where U-Net enabled the generation of synthetic MRIs that closely resembled the ground truth, allowing for more accurate brain structure segmentation. Another key study by Han (2017) applied U-Net to generate synthetic CT images from MR data, showing high accuracy in producing visually consistent outputs with minimal artifacts [23]. These applications demonstrated the versatility of U-Net in learning the mapping between different imaging modalities, making it an attractive choice for cross-modality medical image synthesis.
Despite its success, U-Net does have limitations when applied to image synthesis tasks. The architecture was originally developed for segmentation, not for generating entirely new images, and this can become a challenge. Its reliance on traditional loss functions, such as the L1 loss, can lead to outputs that suffer from blurriness and lack sharp anatomical details. This is because L1 loss minimizes the pixel-wise difference between generated and ground truth images, often averaging pixel intensities, resulting in overly smooth or visually implausible synthetic images [24]. These limitations highlighted the need for more advanced models, such as Generative Adversarial Networks (GANs) [25], which can better capture the distribution of real images and improve visual fidelity.

Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs), introduced by Goodfellow et al. (2014) [25], represent a breakthrough in deep learning for image synthesis by framing the problem as

a game between two neural networks: the generator and the discriminator. The generator network attempts to create realistic synthetic images by learning a mapping from a random noise vector to the desired data distribution. In contrast, the discriminator acts as a binary classifier, distinguishing between real images from the training data and synthetic images generated by the generator. The two networks are trained simultaneously, as shown in Figure 1.5, with the generator improving its outputs to fool the discriminator, while the discriminator becomes better at detecting fake images. Over time, the generator produces increasingly realistic images, and the discriminator struggles to differentiate between real and synthetic data.

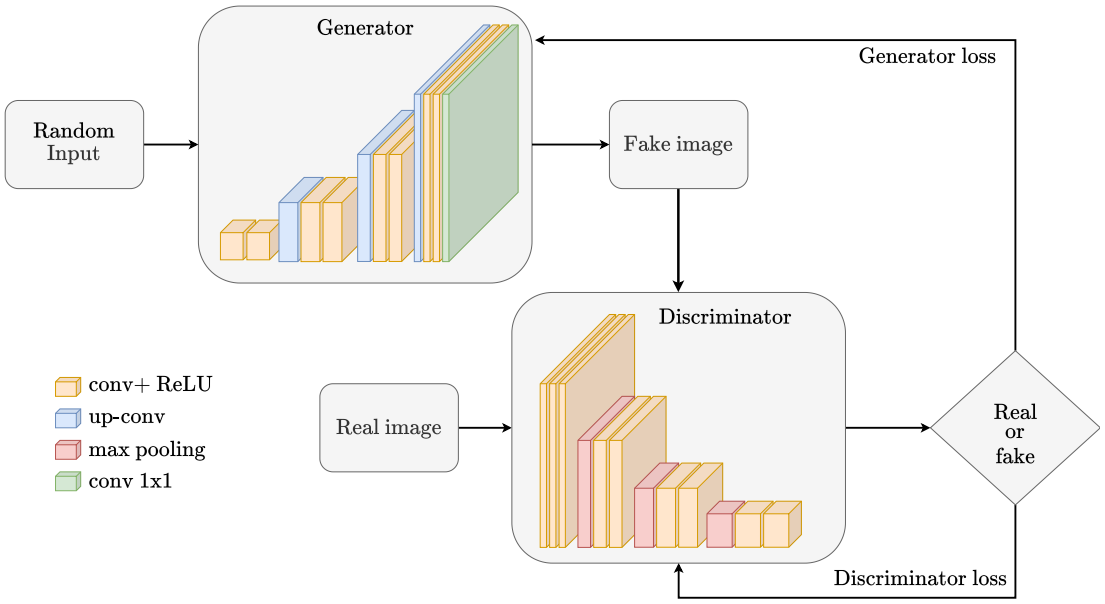


Figure 1.5: Example of GAN architecture for image synthesis from random noise.

The key novelty of GANs lies in their ability to synthesize high-quality images by learning the underlying distribution of the target data, without relying on predefined pixel-based loss functions like L1 or L2, which encourage finding pixel-wise averages of plausible solutions and often result in overly smooth images with poor perceptual quality [24, 26]. Instead, the adversarial loss enables GANs to learn fine details and textures, making them highly effective for tasks such as medical image synthesis, where preserving anatomical details is critical.

GANs have been widely adopted in the medical imaging field due to their strengths in generating high-fidelity images. One notable application is the work by Nie et al. (2017) [27], where GANs were employed to synthesize medical images with contextual awareness, ensuring that the generated images were not only visually plausible but also consistent with the surrounding anatomical structures. This was a crucial development in improving the clinical applicability of synthetic images. Another important example is Han et al. (2018), who applied GANs to generate synthetic brain MR images, demonstrating how the model could create realistic MR images that closely resembled the original data [28]. This approach has proven its value for augmenting datasets, where limited availability of certain modalities or contrast levels can hinder the training of deep learning models.

While GANs offer significant advantages for image synthesis, they also come with challenges, including training instability, mode collapse (where the generator produces a lim-

ited variety of images), and high computational demands [29]. These issues are particularly relevant in medical applications, where generating images that reflect subtle variations in anatomy is crucial for accurate diagnosis. Despite these limitations, the success of GANs in capturing the fine details necessary for medical imaging tasks has made them a vital tool for cross-modality synthesis, segmentation, and augmentation in recent years.

Recent Advances: Diffusion Models

Diffusion models, introduced by Sohl-Dickstein et al. (2015) [30] and further refined in Ho et al. (2020) with *Denoising Diffusion Probabilistic Models* (DDPM) [31], represent a probabilistic alternative to GANs. The core principle of diffusion models is to generate data by reversing a process that gradually adds noise to an image. Starting from a noisy image, the model learns to progressively denoise it, eventually producing high-quality synthetic data. Unlike GANs, which rely on a discriminator to judge the authenticity of generated images, diffusion models learn to generate realistic data by modeling the underlying noise distribution and reversing the process. This approach leads to improved stability during training and fewer artifacts in the generated images.
Diffusion models have gained popularity in medical image synthesis due to their ability to generate high-quality images without the training instability and mode collapse issues commonly associated with GANs. A key advancement in this area is the Latent Diffusion Model (LDM), introduced by Rombach et al. (2021) [32]. LDMs take the process further by operating in a compressed latent space, allowing for more efficient image generation with reduced computational cost compared to the original DDPM formulation. This makes LDMs particularly attractive for high-resolution medical image synthesis, where balancing image quality and computational efficiency is crucial.

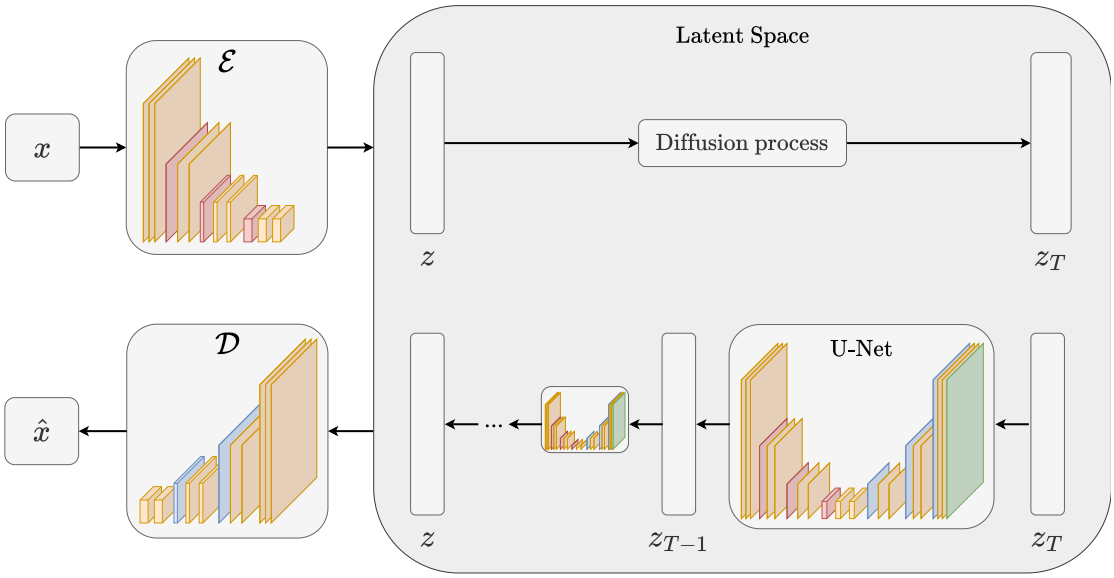


Figure 1.6: Example of latent diffusion model (LDM) architecture for image synthesis.

The process, as illustrated in Figure 1.6, begins with an input image x, which is passed through an encoder $\mathcal{E}$ that compresses the image into a latent representation z. This latent space allows for more efficient manipulation of the image, reducing its complexity

while retaining critical information. Once the image is in the latent space, a diffusion process is applied, during which noise is gradually added to the latent representation until it becomes nearly indistinguishable from random noise z_T. The task of the model is to then reverse this process, starting from z_T, and progressively remove the noise at each time step, represented as z_{T-1}, z_{T-2}, and so on, until the denoised latent representation z is obtained. This iterative denoising is usually carried out by a U-Net, which operates within the latent space and specializes in capturing both local and global features, ensuring the preservation of fine structures throughout the denoising process. Once the final latent variable z is achieved, it is passed through a decoder $\mathcal{D}$, which converts the latent representation back into the high-dimensional image space to generate the final synthetic image $\hat{x}$. The combination of U-Net and decoder ensures that the fine details and anatomical structures of the image are preserved while efficiently handling the noise reversal.

In medical imaging, latent diffusion models have already shown significant promise. Pinaya et al. (2022) applied LDMs to generate synthetic brain MR images, demonstrating that these models can effectively synthesize high-quality medical data that retain critical anatomical structures [33]. Similarly, in Müller-Franzes et al. (2023), a multimodal comparison between GANs and denoising diffusion probabilistic models (DDPMs) was performed for medical image synthesis [34]. In their work, the authors introduced Medfusion, a conditional latent DDPM, and evaluated its performance against GAN-based models on various medical datasets, including fundoscopy, radiographs, and histopathology images. The results showed that Medfusion exceeded GANs in terms of both image diversity and fidelity, with fewer artifacts and better preservation of image details. These findings highlight diffusion models as a promising alternative to GANs for generating diverse and high-quality medical images.

However, despite their strengths, diffusion models are still a relatively new architecture in the field of image synthesis, with several limitations that need to be addressed. One of the main challenges is their long training times and slow inference speed compared to CNN-based methods like GANs. Moreover, diffusion models are often constrained by memory limitations, particularly in 3D image generation, making most current applications restricted to 2D images [35]. As the field progresses, these models are expected to evolve and become more computationally efficient, but for now, they represent a promising direction with room for improvement in their application to large-scale medical datasets.

1.2 Image-to-Image Translation in Medical Imaging

In the previous section, we explored various image synthesis architectures, ranging from traditional approaches to deep learning methods such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), and diffusion models. These architectures enable the synthesis of new images based on a given dataset, allowing for the generation of realistic medical images.

In our specific context, we aim to synthesize one imaging modality from another, specifically, generating computed tomography (CT) images from magnetic resonance (MR) scans. This objective situates our work within the subdomain of image-to-image translation. This field has gained substantial interest, primarily due to the development of conditional GANs (cGANs) [36]. These models extend the adversarial framework of GANs by incorporating an input condition, thereby guiding the image generation process.

Unlike traditional GANs that generate images from random noise, cGANs condition the

generation on a given input image. This conditioning forces the generator to learn a mapping from the input image to the desired output image, effectively translating between domains. The generator in cGANs often adopts a U-shaped architecture, designed to capture both global context and fine-grained details. An example of this architecture is illustrated in Figure 1.7.

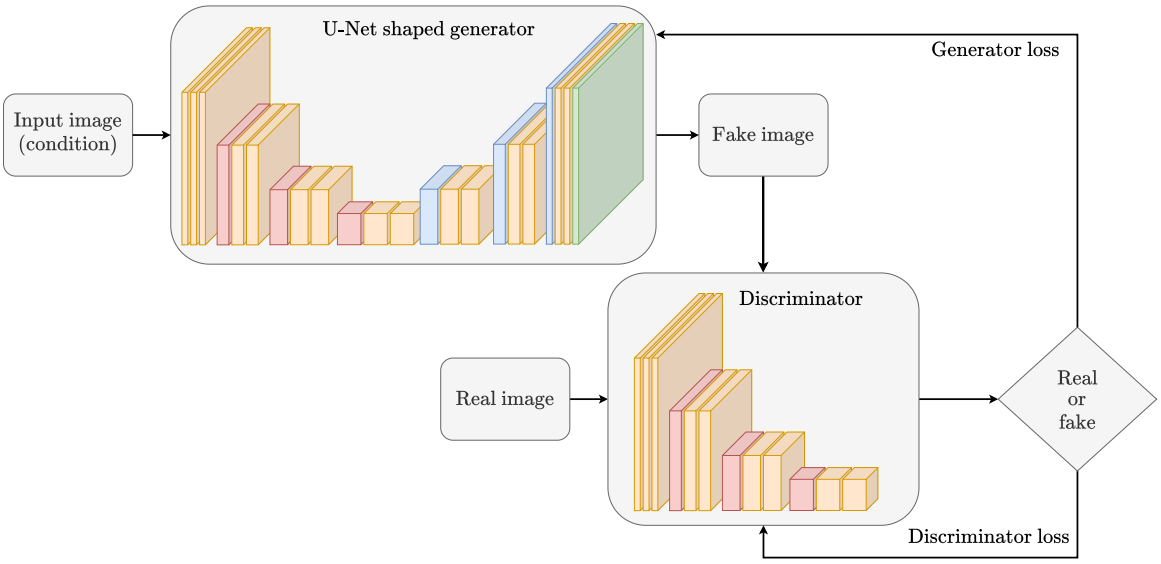


Figure 1.7: Example of paired conditional GAN (cGAN) architecture for image translation.

Conditional GANs can be broadly categorized into two approaches: paired and unpaired methods [37]. Paired methods rely on having matched image pairs for training, where each input image corresponds to a known output. Unpaired methods, on the other hand, are designed to handle scenarios where such paired data is unavailable, making them particularly useful in medical contexts where paired datasets are difficult to obtain. In this section, we will describe existing cGAN architectures, focusing on both paired and unpaired approaches, and highlight successful applications of these methods in medical imaging.

1.2.1 Paired Approaches

Early U-Net based methods

Early attempts at cross-modality image synthesis in medical imaging primarily relied on U-Net-shaped architectures trained with an L1 loss function to learn the direct mapping between input and output images [38]. The U-Net architecture [21], known for its encoder-decoder structure with skip connections, excels at capturing both global context and fine-grained details, making it suitable for medical image translation tasks. Methods utilizing this approach aimed to minimize the pixel-wise differences between the synthesized and target images using the L1 loss:

$$\mathcal{L}_{L1}(G) = \sum_{i=1}^{N} |y - G(x)| \tag{1.1}$$

where G is the generator network, x is the input image, and y is the target image. While this approach could capture the overall structure of the target images, it often led to blurry results due to the limitations of the L1 loss in modeling high-frequency details.

The pix2pix model : revolutionizing image translation

The introduction of the pix2pix model by Isola et al. [39] revolutionized image-to-image translation by integrating a discriminator network into the U-Net architecture, forming a paired conditional GAN (cGAN). This model enhanced the quality of synthesized images by introducing an adversarial loss that encourages the generator to produce outputs indistinguishable from real images, thereby adding a constraint on realism and counterbalancing the tendency of the L1 loss to produce blurred outputs. The adversarial loss in pix2pix is defined as :

$$\mathcal{L}_{GAN}(G, D) = \mathbb{E}_y[logD(y)] + \mathbb{E}_x[log(1 - D(G(x)))] \tag{1.2}$$

where D is the discriminator network that evaluates the authenticity of the generated image $G(x)$ conditioned on the input image x.
In medical imaging, pix2pix has been successfully applied to tasks such as MR to CT synthesis. For instance, Nie et al. [27] employed a pix2pix-based framework to generate CT images from MR scans, facilitating improved radiation therapy planning. Their results demonstrated that incorporating the adversarial loss significantly enhanced the sharpness and structural fidelity of the synthesized images compared to models using only L1 loss.

Enhancements over pix2pix

While pix2pix was initially implemented for 2D images, recent studies have extended its application to 3D medical data. Wang et al. [40] introduced a 3D patch-based approach that adapts the pix2pix framework for volumetric image synthesis. By processing 3D patches instead of full volumes, they managed to capture spatial contextual information effectively while maintaining computational feasibility. This method has also seen success in the recent SynthRAD 2023 challenge [41], where several participants who utilized pix2pix, in both 2D and 3D forms, ranked among the top 10 performers, further demonstrating its versatility and effectiveness in medical imaging [42].

Building upon the foundation of pix2pix, the pix2pixHD model [43] introduced several advancements to enhance the quality of high-resolution image synthesis. One significant improvement is the use of multi-scale discriminators, which evaluate the realism of generated images at multiple spatial scales, allowing for better capture of both global structure and local details.
Additionally, pix2pixHD incorporates a feature matching loss to stabilize training and improve image quality. This loss minimizes the differences between the intermediate feature representations extracted from the discriminator when processing real and generated images:

$$\mathcal{L}_{FM}(G, D) = \mathbb{E}_{x,y} \sum_{i=1}^{T} \frac{1}{N_i} [||D^{(i)}(x, y) - D^{(i)}(x, G(x))||_1] \tag{1.3}$$

where T is the total number of layers, N_i is the number of elements in each layer, and $D^{(i)}$ is the *i*-th layer of the discriminator D.

More recently, the SPADE (Spatially-Adaptive Denormalization) model [44] can be seen as an upgrade to pix2pixHD, introducing a spatially adaptive normalization layers conditioned on the semantic layout of the input image. This innovation allows the generator to preserve spatial and semantic information more effectively during the translation process. By modulating the normalization parameters with spatial information, SPADE enhances the fidelity and realism of the synthesized images, reducing artifacts and improving the representation of fine details.

Addressing misalignment in paired data

Despite the advancements, a significant challenge with paired methods is the necessity for precise pixel-wise correspondence between the input and target images. In medical datasets, achieving such alignment is often difficult due to factors like patient movement, varying acquisition protocols, and differences in anatomical positioning between imaging sessions.
To address the issue of misalignment in paired datasets, Reg-GAN [45] was proposed as a 2D GAN-based method that integrates a registration network into the image synthesis framework. In REG-GAN, a registration network R is placed after the generator G to align the synthesized image with the target image. The model introduces a smoothness term in the loss function to evaluate and minimize the gradient of the deformation field produced by the registration network, ensuring smooth and plausible deformations.
REG-GAN was validated on the BraTS brain tumor dataset [46], specifically focusing on T1 to T2-weighted image synthesis. The results demonstrated that Reg-GAN effectively handled misaligned data, producing synthesized images with improved structural consistency and visual quality compared to traditional paired methods without registration components.
In summary, paired approaches in image-to-image translation have significantly advanced cross-modality synthesis in medical imaging by leveraging direct mappings between aligned datasets. The evolution from basic U-Net models with L1 loss to sophisticated cGAN architectures like pix2pix, pix2pixHD, and SPADE has led to substantial improvements in the realism and structural fidelity of synthesized images. Moreover, innovations like REG-GAN address practical challenges such as data misalignment, enhancing the applicability of these methods in real-world clinical settings.

1.2.2 Unpaired Approaches

Unpaired methods tackle image-to-image translation without the necessity of corresponding image pairs, making them valuable in scenarios where acquiring paired datasets is impractical or costly, a common situation in medical imaging. These methods enable models to learn mappings between domains using unpaired data, thereby broadening the applicability of image translation techniques in clinical settings.

CycleGAN: Bridging Domains Without Paired Data

A trendsetting model in this domain is the CycleGAN proposed by Zhu et al. [47], which introduces a cycle-consistency loss in addition to adversarial losses to facilitate unpaired image-to-image translation. The architecture consists of two generators, G_A and G_B, and two discriminators D_A and D_B. The generators learn mapping between the source

domain A and the target domain B in both directions $G_A : A \to B$ and $G_B : B \to A$, as illustrated in Figure 1.8.

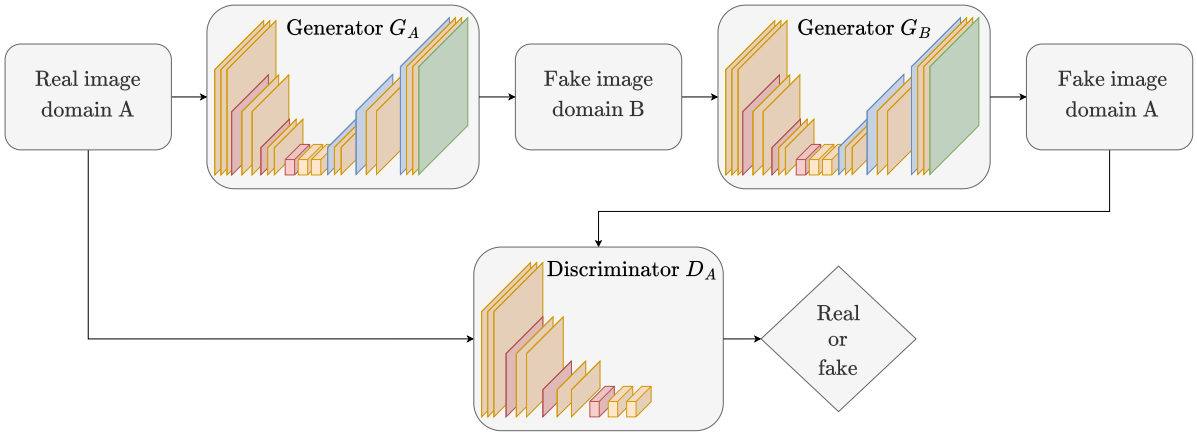


Figure 1.8: Simplified CycleGAN architecture illustrating cycle-consistency for unpaired image translation.

The cycle-consistency loss is defined as:

$$\mathcal{L}_{Cyc}(G_A, G_B) = \mathbb{E}_{x\sim X}[||G_B(G_A(x)) - x||_1] + \mathbb{E}_{y\sim Y}[||G_A(G_B(y)) - y||_1] \tag{1.4}$$

where G and F are the generators for the forward and backward translations, respectively. CycleGAN has been extensively applied in medical imaging for cross-modality translation due to its ability to learn from unpaired data. Authors in [48] employed CycleGAN to generate CT images from MR scans without paired training data. Their work demonstrated that the synthesized CT images retained critical anatomical structures, making them useful for radiation therapy planning where CT images are essential for dose calculations. Wang et al. [40] used CycleGAN to enhance low-dose CT images to standard-dose quality, reducing noise and artifacts while maintaining diagnostic features, thus potentially lowering radiation exposure to patients. Hiasa et al. [49] extended the CycleGAN approach for MR image synthesis from CT scans in orthopedic imaging by introducing a gradient consistency loss. This addition aimed to enhance the accuracy of the generated images, particularly at the boundaries of anatomical structures.

Advancements over CycleGAN

Building upon the foundation laid by CycleGAN, several models have been proposed to address its limitations and enhance performance.
NICE-GAN [50] extends unpaired image translation by reusing the discriminator as an encoder for the target domain. This shared encoder-decoder architecture allows the model to better capture domain-specific features and improves the quality of the generated images. The discriminator's dual role is facilitated through a reconstruction loss, that enhances the structural preservation in synthesized images.
UNIT [51] (Unsupervised Image-to-Image Translation Networks) introduces a shared latent space assumption, where images from different domains can be mapped to a common latent representation. The model comprises two variational autoencoders (VAE) for the source and target domains, sharing weights in their latent space encoders and decoders.

CUT [52] (Contrastive Unpaired Translation) introduces a contrastive learning framework for unpaired image translation, eliminating the need for a cycle-consistency loss. Instead, it maximizes mutual information between corresponding patches in the input and output images using a contrastive loss. CUT encourages the generator to preserve content while allowing for style changes between domains.

Challenges in unpaired methods

Despite the advancements of unpaired methods like CycleGAN in image-to-image translation, these approaches present significant challenges in medical imaging. A key concern is their propensity to hallucinate features, introducing or omitting critical anatomical details, due to their reliance on distribution matching losses to statistically align source and target domains. This reliance means that if certain classes or features (such as specific diseases or anomalies) are underrepresented or overrepresented in the target domain, the model may inadvertently alter these features in the translated images, potentially leading to misdiagnosis.
Recent studies highlight this issue [53], emphasizing that while unpaired models produce high-quality images without the need for paired data, they may not faithfully preserve all known and unknown class labels. The absence of explicit guidance in unpaired methods means there is no guarantee that clinically relevant features are accurately translated. Therefore, it is essential to approach the use of unpaired image translation models with caution in medical contexts. Translated images from these models should not be used for direct clinical interpretation without thorough validation, as they may introduce uncertainties and pose risks to patient safety.

1.3 Evaluation Metrics for Synthesized Images

After presenting various architectures for image synthesis and translation, including CNNs, U-Net, GANs, and diffusion models, it is crucial to have tools to evaluate the quality of the synthesized images. Image quality assessment plays a pivotal role in determining the effectiveness of these models, particularly in medical imaging, where precision is paramount. In this section, we will introduce key evaluation metrics commonly used to assess image quality, including both generic and GAN-specific metrics. Furthermore, we will discuss the challenges of evaluating synthesized images in medical contexts, where clinical relevance is as important as signal quality.

1.3.1 Intensity-based Metrics

Pairwise comparison is one of the most common techniques used to evaluate medical image synthesis. It involves comparing the synthesized image to a reference image (often referred to as the ground truth), allowing for a direct assessment of the quality of the generated image. Several generic metrics are used in this process, including *Mean Absolute Error (MAE)*, *Peak Signal-to-Noise Ratio (PSNR)*, and *Structural Similarity Index Measure (SSIM)*. Each of these metrics, while widely adopted, come with their own strengths and limitations when applied to medical image synthesis.

Mean Absolute Error (MAE) One of the simplest and most frequently used metrics, MAE calculates the average absolute difference between pixel intensities in the synthesized

and reference images, defined as :

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{1.5}$$

where y_i and x_i are the intensity values of the i-th pixel in the synthesized and real image, respectively, and n is the total number of pixel pairs. MAE is particularly useful in tasks such as CT synthesis from MRI for attenuation correction, where quantitative accuracy is essential for accurate dose planning.

Peak signal-to-noise ratio (PSNR) Derived from the Mean Squared Error (MSE), PSNR is a widely used metric in image processing. It compares the maximum possible pixel intensity in the image to the error between the synthesized and ground truth images :

$$PSNR = 10\log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{1.6}$$

where MAX represents the maximum intensity value in the image and MSE is the mean squared error. PSNR is effective for detecting global intensity differences but has limitations in capturing structural differences, which are often critical in medical image synthesis. Higher PSNR values indicate better quality images.

Structural Similarity Index Measure (SSIM) While metrics like MAE and PSNR are simple to compute, their ability to accurately assess visual quality is limited, especially when dealing with perceptual differences. To address this, Wang et al. (2004) [54] introduced the Structural Similarity Index Measure (SSIM), which exploits known characteristics of the human visual system by comparing local patterns of pixel intensities, normalized for luminance and contrast. SSIM has since become widely adopted in the image synthesis community for its ability to assess the structural integrity of images.
SSIM is computed by evaluating three components, luminance, contrast, and structure, between the synthesized image and the reference image:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1.7}$$

where x and y represent the synthesized and reference images, μ_x and μ_y are the mean values of the pixel intensities in the images, σ_x^2 and σ_y^2 are the variances, and σ_{xy} is the covariance between the two images. The constants C_1 and C_2 are small values included to avoid instability when the denominator is close to zero. The higher the SSIM value, the better the perceived quality of the synthesized image, particularly in terms of structural consistency.
SSIM is often preferred over MAE and PSNR because it evaluates structural changes, making it more aligned with human perception of image quality. It has been used extensively in medical image synthesis, where maintaining structural integrity is critical, for example in tasks such as MR to CT synthesis and image super-resolution.
Despite its advantages, SSIM is not without its limitations. Studies have shown that, while SSIM and PSNR are correlated, they have different sensitivities to image degradations. For instance, Mason et al. (2020) [55] found that neither SSIM nor PSNR perfectly correlated with radiologists' subjective assessments of MR image quality. These findings suggest that, while SSIM is a useful tool for image quality evaluation, it may not fully

capture the clinical relevance of synthesized images. Nonetheless, it remains a de-facto standard metric for pairwise image comparison and is often used in combination with other metrics such as MAE and PSNR to provide a more comprehensive evaluation of synthesized images.

1.3.2 Distribution-based Metrics

As image synthesis techniques have evolved, particularly with the introduction of Generative Adversarial Networks (GANs), traditional image quality assessment metrics like MAE, PSNR, and SSIM often fall short in evaluating the nuanced differences between real and generated images. GAN-specific metrics, which focus on comparing probability distributions rather than just pixel-level differences, have been developed to assess the fidelity and diversity of synthesized images. In this section, we will focus on three key metrics: *Mutual Information (MI)*, *Maximum Mean Discrepancy (MMD)* , and the *Fréchet Inception Distance (FID)*.

Mutual Information (MI) Mutual Information (MI) quantifies the amount of shared information between two images, measuring how much knowing the intensity of one image reduces the uncertainty of the intensity in the other. It is particularly useful when direct pixel-wise correspondence between real and synthesized images is not available, for example unpaired medical image synthesis tasks using CycleGAN [48]. MI is defined as:

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{1.8}$$

where $H(X)$ and $H(Y)$ are the entropies of images X and Y, and $H(X,Y)$ is their joint entropy. Entropy H measures the amount of uncertainty or variability in the intensity distribution of an image, and it is calculated as:

$$H(I) = -\sum_{a} p_I(a) \log p_I(a) \tag{1.9}$$

where $p_I(a)$ is the probability distribution of pixel intensities in image I. A higher MI value indicates greater similarity between the images.

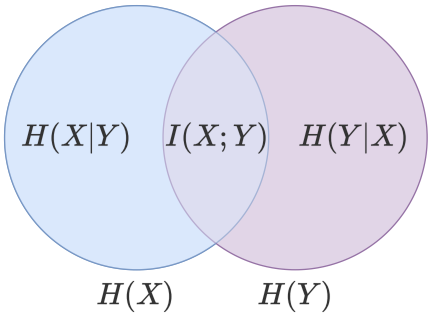


Figure 1.9: Representation of mutual information (MI) between two variables.

Figure 1.9 illustrates Mutual Information (MI) for two correlated variables, X and Y. The left circle (blue) represents the entropy $H(X)$, and the right circle (purple) represents the entropy $H(Y)$. The overlapping area corresponds to the mutual information $I(X;Y)$, indicating the shared information between X and Y. The non-overlapping parts depict the conditional entropies: $H(X \mid Y)$ and $H(Y \mid X)$.

Maximum Mean Discrepancy (MMD) [56] A statistical measure used to assess the difference between two probability distributions based on samples drawn from each. In the context of image synthesis, MMD is used to compare the distributions of real and synthesized images. The goal is to determine how well the generated images match the distribution of the real ones. MMD is calculated using kernel functions to measure the discrepancy between two distributions:

$$MMD^2(Pr, Pq) = E[k(x_r, x_r')] - 2E[k(x_r, x_q)] + E[k(x_q, x_q')] \tag{1.10}$$

where Pr and Pq are the distributions of real and synthesized images, respectively, and k is a kernel function (e.g., a Gaussian kernel). A lower MMD value suggests that the two distributions are more similar. MMD is useful in GANs for assessing mode collapse, where the generator fails to capture the diversity of the data. This metric has been applied to evaluate the quality of generated data, such as in medical image synthesis for tasks like multi-modal MRI generation.

Fréchet Inception Distance (FID) [57] A popular GAN-specific metric, which evaluates the similarity between distributions of real and generated images by comparing their feature representations. Unlike pixel-based measures, FID leverages a pre-trained network (Inception v3) to extract feature vectors from both real and generated images. The distance between the distributions is then calculated as:

$$FID = ||\mu_1 - \mu_2||^2 + Tr(C_1 + C_2 - 2 \cdot \sqrt{C_1 \cdot C_2}) \tag{1.11}$$

where μ_1 and μ_2 are the means of the feature vectors, and C_1 and C_2 are their covariance matrices. FID measures both the fidelity and diversity of the generated images, with a lower FID value indicating better quality. This metric is widely used in medical image synthesis tasks, where it helps evaluate how closely the generated images resemble real medical images in a feature space.

1.3.3 Challenges in Evaluating Medical Image Synthesis

While intensity-based and distribution-based metrics are widely used in evaluating medical image synthesis, they fall short in capturing the clinical relevance and fine anatomical details that are critical in medical imaging. These challenges arise from the inherent limitations of these metrics, which focus either on pixel-wise intensity differences or global distributional similarities, often missing clinically significant features and localized image details.

Limitations of current metrics

Intensity-based metrics primarily focus on pixel-level differences between the synthesized and reference images. While these metrics are effective at measuring numerical accuracy, they do not provide any perceptual insights. This means that an image can have a low MAE or high PSNR but still fail to capture critical anatomical features that are essential in clinical diagnosis. For instance, small lesions or fine structures like airways or blood vessels may be lost or distorted, yet these metrics would not reflect such deficiencies. As pointed out by Xiang et al. (2018) [58], intensity-based metrics do not fully capture the preservation of clinically relevant features. Thus, these metrics may give a misleading impression of quality in cases where the reconstructed image is numerically similar but

visually and anatomically inaccurate. Moreover, recent studies have shown that similarity scores tend to favor blurry images, especially SSIM [59].
On the other hand, distribution-based metrics like the Fréchet Inception Distance (FID) and Maximum Mean Discrepancy (MMD) provide a broader statistical assessment by comparing the distributions of real and generated images. These metrics can offer more perceptual insights than intensity-based metrics because they compare feature representations rather than pixel values. However, they still fall short in medical imaging, as they primarily capture global image statistics rather than focusing on the medical significance of local structures. Recent studies have highlighted inherent limitations of these metrics: Binkowski et al. (2018) [60] show that FID is biased regarding sample size and the choice of model, while Chong and Forsyth (2020) [61] demonstrate that the FID score decreases with increasing sample size, leading to inconsistent evaluations. As described by Cohen et al. (2018) [53], distribution-based metrics may fail to reflect the accuracy of pathological features in synthesized medical images, leading to unrealistic or hallucinated features not present in real images. This issue highlights the challenge of using these metrics to evaluate medical images, where accuracy in specific regions (e.g., lesions, tumors, or organs) is far more important than a global match to real data.

Need for Local and Task-specific Metrics

Both intensity-based and distribution-based metrics are global in nature, meaning they assess the entire image rather than focusing on specific regions of interest. In medical imaging, the importance of accurate synthesis varies across different regions, as discussed by Armanious et al. (2020) [62]. For example, the precision required for regions containing critical structures like the heart or brain may differ from less important areas. These metrics fail to account for such context-dependent importance, resulting in evaluations that may not reflect the clinical significance of the synthesized image. Figure 1.10 provides an overview of different evaluation metrics used in medical image synthesis.

To address these limitations, researchers have advocated for the development of task-specific metrics that are aligned with clinical objectives. For example, a study focused on the synthesis of medical images for segmentation should evaluate the performance of the synthesized images based on their utility in the segmentation task [27, 59]. Similarly, classification accuracy or lesion detection rates can provide a more clinically meaningful evaluation of the synthesized images.

The Role of Expert Evaluation

While automated metrics are convenient, they often do not align with clinical perception. There is often a disconnect between statistical measures of image quality and how clinical experts assess the usefulness of an image. However, relying solely on clinical experts for evaluation is time-consuming, labor-intensive, and not always feasible, especially in large-scale studies.
In summary, while intensity-based and distribution-based metrics provide useful tools for initial image quality assessments, they are insufficient for fully evaluating the clinical utility of synthesized medical images. Moving forward, task-specific evaluation strategies and the incorporation of expert feedback are essential to improve the assessment of medical image synthesis in real-world applications.

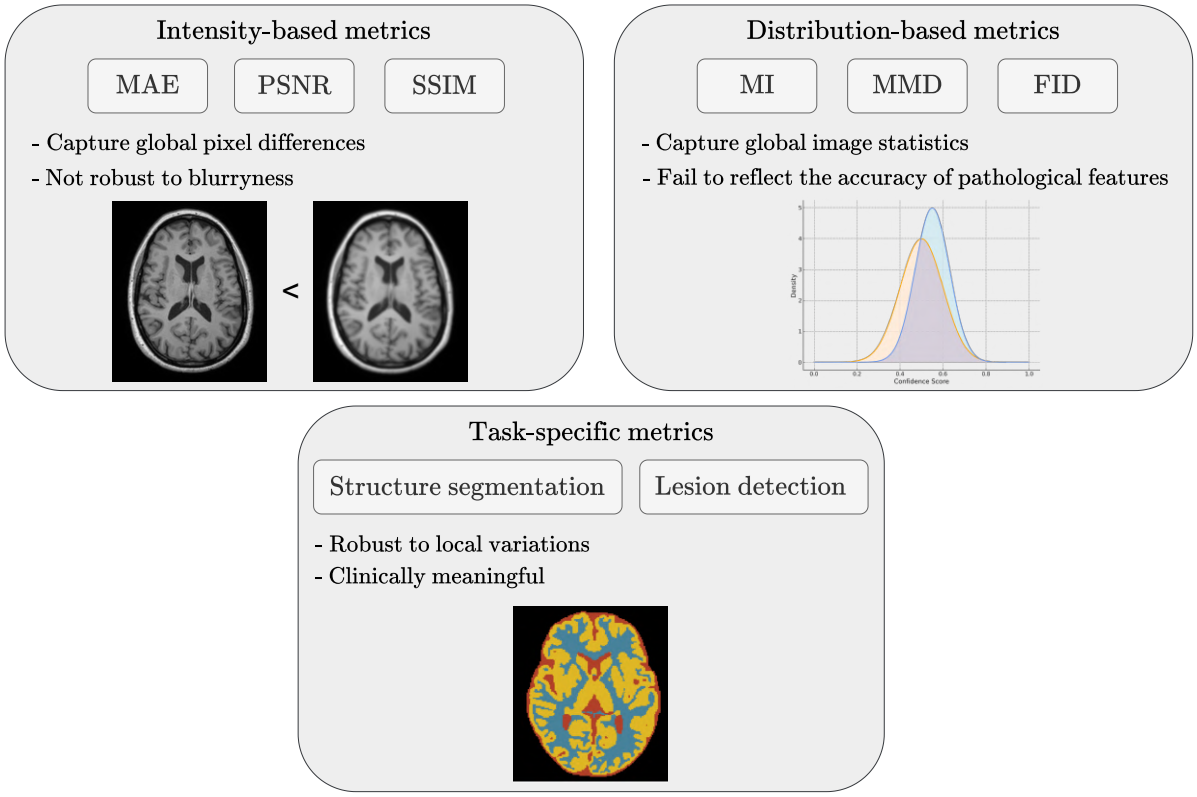


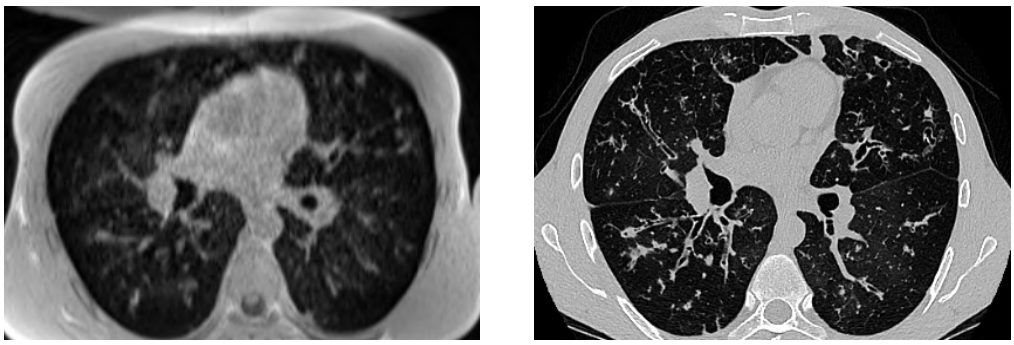
Figure 1.10: Overview of evaluation metrics for medical image synthesis.

1.4 Limitations in Thoracic Imaging

In this section, we introduce key limitations and challenges in applying synthetic architectures within our clinical context: thoracic imaging. Specifically, we aim to generate synthetic lung CT images from UTE MR, a relatively novel modality in thoracic imaging. While cross-modality synthesis has shown success in areas like brain [48], pelvis [63], and aorta [64], there are unique difficulties associated with applying these methods to the lung region. This section will outline the challenges related to the novelty of UTE MR, data acquisition, image quality discrepancies, and the anatomical complexity of thoracic structures.

1.4.1 Novelty and Dataset Limitations

One of the primary challenges in applying MR to CT synthesis to thoracic imaging is the lack of precedent. To date, no previous study has attempted to generate synthetic lung CT images from UTE MR scans. UTE MR is still relatively new in its application to thoracic imaging, and there is limited data available for this task [65].
A critical requirement for cross-modality translation is a paired dataset consisting of corresponding CT and MR images. However, acquiring such a dataset for lung imaging is particularly difficult. Due to the differences in imaging modalities, coupled with the relatively recent introduction of UTE MR in clinical practice, publicly available paired datasets do not exist in the current state of the art. This lack of accessible data presents a significant obstacle to advancing research in this area, as robust datasets are essential for training deep learning models [66]



(a) UTE MR (b) CT

Figure 1.11: Comparison of imaging quality between lung UTE MR and CT. The UTE MR image demonstrates lower spatial resolution, blurriness, and artifacts, whereas the CT image provides higher resolution with clearer visualization of anatomical structures.

1.4.2 Image Quality Challenges

Beyond limitations in data availability, there are inherent challenges in dealing with the quality discrepancies between the two modalities. CT is the gold standard for thoracic imaging, providing high spatial resolution, sharp images with low noise, and clear visualization of lung structures [67]. In contrast, UTE MR, while a promising emerging modality, suffers from lower spatial resolution, higher levels of noise, and artifacts. Blurriness and motion-related distortions are common in UTE MR images, particularly in the lung region, which is susceptible to respiratory motion and has relatively weak proton density [68].
The differences in image quality between these two modalities are critical when considering MR to CT synthesis. CT scans offer clear, well-defined structures, making it easier to identify small anatomical features. UTE MR, on the other hand, presents challenges in interpreting lesions or fine structures due to artifacts and blurriness, as illustrated in Figure 1.11. While the MRI signal can detect lesions and assess fine structures, it often lacks the sharpness needed for precise analysis, complicating its application for synthetic tasks.
In the context of MR to CT translation, these difficulties must be addressed to ensure the synthesized images are both anatomically accurate and of high-quality. Moreover, adaptations to the standard architectures may be necessary to avoid replicating the blurriness and artifacts present in the input MR images. Drawing inspiration from methods used in super-resolution [69], our task involves not only translating between modalities but also enhancing the resolution and quality of the synthesized CT images to match that of the original CT scan.

1.4.3 Anatomical Complexity of Thoracic Structures

Another significant challenge in thoracic imaging is the complexity of the anatomical structures in the lungs. The thoracic region contains numerous small and fine structures, such as blood vessels, airways, and bronchi, which are essential for diagnosing lung diseases and assessing lesion severity. Radiologists often rely on the precise visualization of these structures to detect subtle pathological changes or track the progression of diseases like cystic fibrosis or lung cancer [68].
These fine structures present a particular challenge in MR to CT synthesis, as standard

architectures may not be able to adequately preserve or reconstruct them. While general image synthesis tasks might focus on broader image regions, thoracic imaging requires extreme precision in local areas where fine anatomical features are critical. Therefore, attention must be paid to ensure that the synthesis process accurately captures and maintains the integrity of these structures. Any loss of detail in these regions could result in the misidentification of lesions or other clinical inaccuracies, rendering the synthetic images less useful for clinical applications.

In summary, while cross-modality synthesis has been proven effective in other areas of medical imaging, thoracic imaging presents unique challenges. These include the novelty of UTE MR, the lack of paired datasets, discrepancies in image quality between MR and CT, and the complex anatomical requirements of the lung region. Addressing these challenges will be crucial for advancing the application of MR to CT synthesis in thoracic imaging.

Chapter Conclusion

This chapter laid the foundation for understanding image synthesis, particularly in the context of medical imaging and the task of MR to CT translation. We began by exploring the evolution of image synthesis methods, highlighting traditional approaches, followed by the introduction of machine learning and deep learning models. Notably, architectures like U-Net and Generative Adversarial Networks (GANs) have revolutionized this field. U-Net's encoder-decoder architecture, initially designed for segmentation, has been widely applied in medical image synthesis. GANs, and more specifically conditional GANs (cGANs), have enabled successful cross-modality translations, with notable implementations in brain and pelvis imaging. However, the transition of these methods to thoracic imaging poses several challenges. Unlike other regions, the lungs present difficulties due to the novelty of UTE MR, the absence of a large, publicly available paired dataset, and the inherent discrepancies in image quality between MR and CT.

We also discussed the evaluation of synthesized images, noting that existing metrics fall short in capturing clinically significant details. Intensity-based metrics like MAE and PSNR, while useful for overall image comparison, do not account for the perceptual quality or anatomical precision necessary in medical imaging. Similarly, distribution-based metrics, such as MMD and FID, capture global image statistics but fail to reflect the specific requirements of thoracic imaging, where small structures such as vessels and airways are crucial for clinical analysis. This underscores the need for more localized, task-specific metrics that prioritize anatomical accuracy, as well as the importance of expert evaluation to validate the clinical relevance of synthesized images.

Looking ahead, the next chapter will focus on the feasibility of lung MR to CT synthesis, building on the limitations discussed in this chapter. We will begin by investigating the performance of 2D architectures for this specific task, followed by the optimization of data preprocessing techniques such as resampling, registration, and intensity normalization. Finally, a clinical assessment will be conducted in collaboration with expert radiologists, evaluating the ability of the synthesized images to represent critical anatomical structures in the lungs, such as bronchi and vessels, which are essential for accurate diagnosis and disease assessment.

Chapter 2

MR to CT Synthesis using 2D GANs : Feasibility and Preprocessing Optimization

The methods discussed in this chapter are related with the following publications :

[1] **Arthur Longuefosse**, Julien Raoult, Ilyes Benlala, Baudouin Denis de Senneville, Thomas Benkert, Julie Macey, Stéphanie Bui, Patrick Berger, Gilbert Ferretti, Jean-Yves Gaubert, Renan Liberge, Antoine Hutt, Baptiste Morel, François Laurent, Fabien Baldacci, Gaël Dournes. "Generating high-resolution synthetic CT from lung MRI with ultrashort echo times: initial evaluation in cystic fibrosis". *Radiology*, 308 (2023).

[3] **Arthur Longuefosse**, Baudouin Denis de Senneville, Gaël Dournes, Ilyes Benlala, Pascal Desbarats, Fabien Baldacci. "Lung CT Synthesis Using GANs with Conditional Normalization on Registered Ultrashort Echo-Time MRI". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2023).*

[4] **Arthur Longuefosse**, Baudouin Denis de Senneville, Gaël Dournes, Ilyes Benlala, Pascal Desbarats, Fabien Baldacci. "MR to CT synthesis using GANs: a practical guide applied to thoracic imaging". In *International Conference on Information Visualization Theory and Applications, IVAPP (2023).*

In the previous chapter, the fundamentals of image synthesis and medical image translation were presented, with a particular focus on both traditional and deep learning approaches, evaluation metrics for synthesized images, and the challenges encountered in the application to lung imaging. To address these limitations, this chapter focuses on the feasibility of MR to CT synthesis in lung imaging. Our objective is to explore and compare various supervised and unsupervised generative models, assessing their ability to synthesize high-quality CT images from MR data. This evaluation includes an in-depth analysis of the impact of different preprocessing strategies, such as resampling, registration and intensity normalization, on the synthesis results. Finally, this chapter presents a clinical assessment conducted by expert radiologists. This evaluation will focus on the ability of the synthesized images to accurately represent critical lung structures and lesions, such as vessels and bronchi, which are essential for diagnosing lung diseases like cystic fibrosis.

2.1 Feasibility of MR to CT Lung Translation

The first study of this thesis is dedicated to exploring the feasibility of MR to CT translation in lung imaging, a challenging task that has not been previously addressed in this specific anatomical region. While MR to CT synthesis has been investigated in areas such as brain [48, 63], pelvis [63], and aorta [64], translating these methods to the lungs presents unique difficulties due to the complex and dynamic nature of pulmonary structures. These challenges include respiratory motion, which can lead to misalignment between the modalities, and the inherent differences between ultrashort echo-time (UTE) MR and CT scans. UTE MR imaging, while beneficial for visualizing lung tissues, often produces lower signal intensity, increased blurriness, and lower resolution compared to CT [68], further complicating the synthesis process.
To thoroughly assess the feasibility, this section will begin with an overview of the generative algorithms and models employed, particularly focusing on 2D GAN architectures, which offer a balance between computational efficiency and accuracy. Following this, we will introduce the MR/CT lung dataset acquired in the Hospital Center of Bordeaux used in our experiments, detailing the steps necessary to treat it as a paired dataset, a crucial consideration for the application of supervised learning methods. Finally, we will review and compare the performance of various state-of-the-art methods in synthesizing lung CT images from UTE MR data, providing insights into their strengths and limitations.

2.1.1 Exploring 2D GAN Architectures

As detailled in the previous chapter, Generative adversarial networks (GANs) have emerged as powerful tools for cross-modality translation tasks, particularly in the medical imaging domain. Unlike traditional convolutional neural networks (CNNs), which focus solely on learning the mapping between input and output data, GANs introduce a discriminator component that evaluates the quality and realism of the generated images. This adversarial framework helps to produce outputs that are not only accurate but also visually convincing.
For this initial study, we focus on 2D GAN architectures as a first step in assessing the feasibility of MR to CT lung translation. The decision to use 2D architectures is driven by several factors: they are easier to train compared to their 3D counterparts, require less data, and have significantly lower computational costs. Furthermore, 2D architectures enable the direct use of well-established natural image-to-image translation models, which are predominantly based on 2D data. While 3D GANs offer the advantage of capturing volumetric context, their excessive computational demands and the need for larger datasets make them less suitable for initial feasibility studies.
Both unpaired and paired GAN approaches for MR to CT lung translation are explored in this section. Unpaired GANs, such as CycleGAN [47], allow us to perform cross-modality translation without requiring direct pixel-to-pixel correspondence between the MR and CT images. This flexibility is particularly advantageous when working with datasets that lack precise alignment between modalities, as is the case with our current dataset. The differences in acquisition properties between MR and CT, such as variations in resolution, field of view, and respiratory phase synchronization, mean that direct correspondence between images is not readily available, making unpaired methods a natural starting point. However, while unpaired GANs provide a valuable approach in the absence of aligned data, paired GAN methods, such as pix2pix [39], are generally able to produce more

accurate and detailed translations when paired datasets are available. These supervised approaches rely on direct pixel-to-pixel correspondence between input and output images, which allows the model to learn a precise mapping between the MR and CT modalities. In order to utilize these paired methods with our dataset, a preprocessing step of cross-modality registration is thus required. Figure 2.1 illustrates the pipeline for the study.

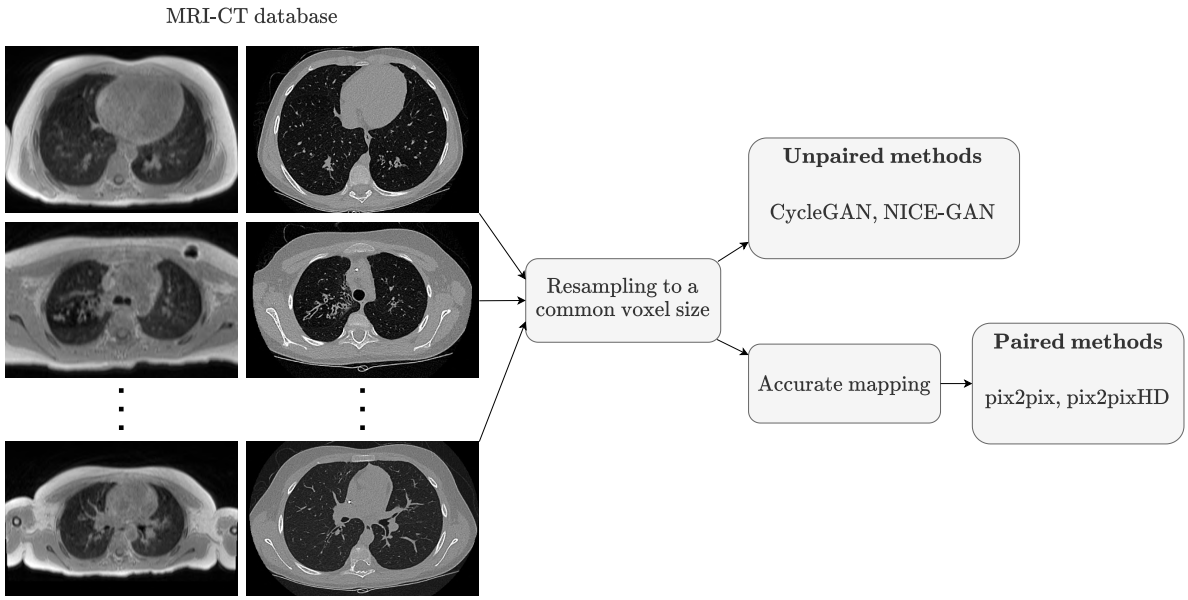


Figure 2.1: Pipeline illustrating the preprocessing steps for paired and unpaired GAN methods.

Unsupervised Approaches In this study, we utilize the well-known CycleGAN model [47], a fundamental method in unpaired image-to-image translation tasks. CycleGAN has already been proven effective in medical cross-modality translation, making it a suitable choice for our dataset, where direct pixel-to-pixel correspondence between MR and CT images is not available.
In addition, we explore NICE-GAN [50], a refinement of the CycleGAN approach. NICE-GAN enhances the traditional CycleGAN framework by incorporating a novel use of the discriminator, which also functions as an encoder for target domain images. This refinement aims to improve structural preservation and limit the excessive deformation that can be produced by CycleGAN.

Supervised Approaches For the paired approaches, we employ pix2pix [39], a widely recognized model for paired image-to-image translation in both natural and medical images.
We also evaluate the effectiveness of pix2pixHD [43], an enhanced version of pix2pix designed to generate higher-resolution images with greater detail. Pix2pixHD improves upon the original pix2pix by incorporating multi-scale discriminators and feature-matching losses, which help to produce sharper and more realistic outputs.

2.1.2 Dataset Insights

The dataset used in this study consists of paired ultrashort echo time (UTE) magnetic resonance (MR) and computed tomography (CT) thoracic images from 110 patients diagnosed with cystic fibrosis. These images were collected as part of their annual follow-up between 2018 and 2022 at the University Hospital of Bordeaux. Both imaging modalities were acquired on the same day for each patient, ensuring temporal consistency and allowing for direct comparison between the MR and CT scans.

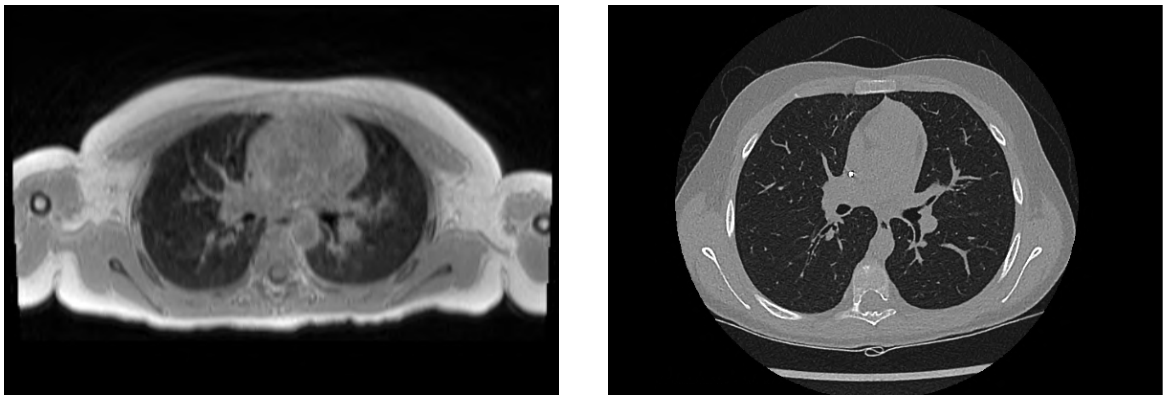
CT Acquisition CT scans were performed using either a Siemens SOMATOM Force or a GE Medical Systems Revolution CT scanner with SAFIRE iterative reconstruction. Patients were instructed to hold their breath at functional residual capacity (end normal expiration) during the acquisition to standardize lung volume across patients. The dose-length product for these scans ranged from 13 to 260 mGy.cm, with a computed tomography dose index volume between 0.6 and 7.5 mGy. The CT images were acquired with a voxel size of $0.6 \times 0.6 \times 0.6$ mm^3, and a median image resolution of $500 \times 512 \times 512$.

MR Acquisition UTE MR images were acquired using the SpiralVibe sequence [70] on a SIEMENS Aera scanner. The key sequence parameters were a repetition time of 4.1 ms, an echo time of 0.07 ms, and a flip angle of 5°. The total acquisition time ranged from 6 to 8 minutes, utilizing prospective respiratory gating at the end of normal expiration to minimize motion artifacts. No contrast agents or sedation were used during the MRI exams. The MR images were acquired with an isotropic voxel dimension of $1 \times 1 \times 1$ mm^3, and a median image resolution of $340 \times 300 \times 340$.

Dataset Considerations A critical consideration in this dataset is the differences in resolution, voxel spacing, and field of view between the CT and MR images. Additionally, while efforts were made to synchronize the acquisition of both modalities by performing them on the same day, there are inherent differences in how each modality captures respiratory motion.
CT scans were performed with the patients holding their breath at functional residual capacity (end normal expiration). This method standardizes lung volume across patients and minimizes motion artifacts during the brief scan time. In contrast, MR acquisitions, which require a longer scan time, were performed during free breathing. The use of prospective respiratory gating helps to minimize motion artifacts, but it cannot fully eliminate variations caused by natural breathing cycles during the MR scan.
Furthermore, the positioning of the patient's arms differs between the two modalities: in CT scans, the arms are positioned above the head to avoid interference, while in MR scans, the arms are placed alongside the body. These differences result in variations in the anatomical content of the images, as depicted in Figure 2.2.

2.1.3 Ensuring Dataset Pairing

To effectively utilize paired GAN methods for MR to CT synthesis, it is essential to establish a voxel-to-voxel correspondence between the modalities. Our dataset consists of MR and CT volumes acquired from the same patients on the same day, making it a strong candidate for paired dataset approaches. However, to fully utilize this potential,



(a) UTE MR (b) CT

Figure 2.2: Comparison between UTE MR and CT images from the same patient in axial view. The CT scan shows larger, more extended lungs due to full-inspiration breath-hold, while the MR scan captures smaller lungs during free breathing. In the MR image, the arms are visible alongside the body, while in the CT image, only the center of the body is present.

several critical preprocessing steps are necessary to ensure that the data is aligned and standardized across both modalities.
Firstly, the MR and CT images must be resampled to a common voxel size to place both modalities on the same spatial grid. This involves interpolation to standardize the resolution and spacing of the images. Once this step is complete, the modalities need to be registered using robust registration algorithms to correct for any misalignments and ensure that corresponding anatomical structures in both images are precisely aligned.
After registration, it is also necessary to define a common field of view (FOV) between the modalities, as MR and CT images are often acquired with different parameters. By standardizing the FOV, we can ensure that the same anatomical regions are covered in both images, facilitating more accurate pairing, as illustrated in Figure 2.3.

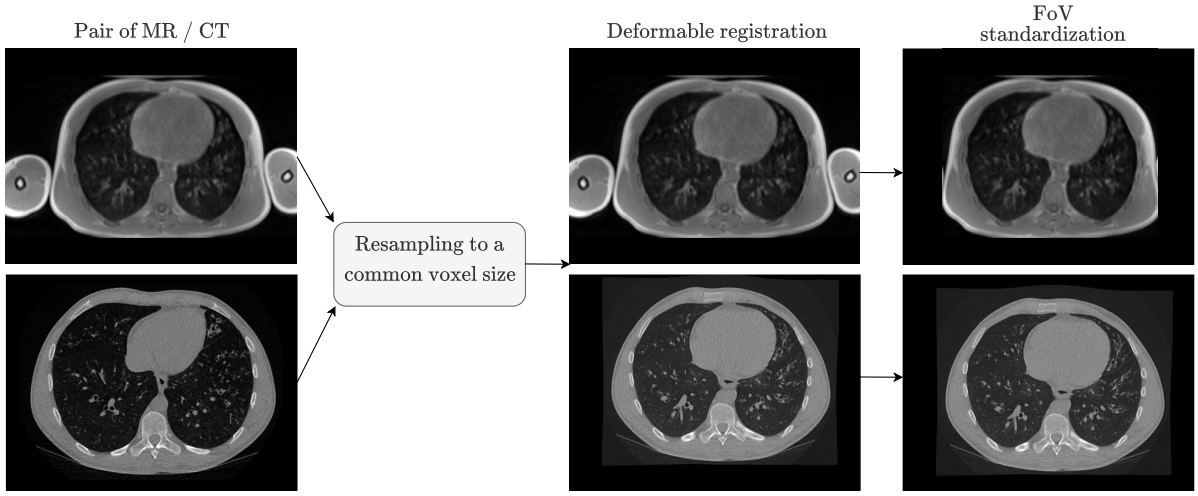


Figure 2.3: Pipeline illustrating the registration steps to ensure dataset pairing.

Finally, intensity normalization is crucial to bring the intensity values of the MR and CT images into a comparable range [71]. Given the fundamental differences between the two modalities, appropriate preprocessing of intensities ensures that both images are prepared

for the paired GAN methods.
In this section, we will detail each of these preprocessing steps, including resampling, registration, field of view alignment, and intensity normalization, to establish a fully paired dataset ready for advanced image synthesis techniques.

Resampling to a Common Grid

The first step in preparing the dataset for paired GAN methods is to resample the images from both modalities onto a common grid with a similar voxel size. Traditionally, the common voxel size is defined by the fixed image [72]. Since our study aims to develop a method based solely on MR data, with MR as the primary input for our synthesis models, the CT images need to be registered onto the MRI. Therefore, the voxel size of the MRI, i.e. $1 \times 1 \times 1$ mm^3, will serve as the reference for resampling.
Resampling ensures that both modalities are standardized in terms of resolution and spatial dimensions, which is crucial for accurate alignment in subsequent registration steps. This process can be accomplished using various interpolation methods, such as linear interpolation, nearest neighbor, or B-spline interpolation. Among these, B-spline interpolation is particularly well-suited for medical imaging due to its ability to produce smooth and accurate results while minimizing artifacts [73].
Using tools such as SimpleITK [74], B-spline interpolation can effectively handle the resampling process, ensuring that the quality of the images is maintained for the next stages of preprocessing, ultimately facilitating more precise registration and pairing of the dataset.

Conventional Image Registration Approaches

After resampling the images to a common grid, the next crucial step is to align the modalities accurately. This is where image registration comes into play, ensuring that corresponding anatomical structures in both the MR and CT images are properly aligned. Image registration is a critical process in medical imaging, enabling the alignment of images from different modalities or time points to facilitate accurate analysis. Various techniques have been developed over the years, ranging from traditional optimization-based methods to modern deep learning approaches. In this section, we review some of the most widely used traditional approaches for image registration.

MIND (Mutual Information-based Normalization of Images for Deformation) [75] The MIND method transforms input images into a vectorial representation, associating each voxel with a descriptor that reflects its relationship with surrounding voxels. By focusing on structural information rather than intensity, MIND provides a robust registration method, particularly for multi-modal images. The algorithm has shown improved performance over other popular methods in various test cases. MIND employs the sum of squared differences (SSD) to measure similarity between descriptors, allowing for fast optimization. However, the method is sensitive to parameters such as noise variance, patch size, and the number of voxels used in descriptor computation. Additionally, it may struggle with low contrast images and situations where the structural information between images differs significantly, such as transient structures or changes in pathology within the field of view.

EVolution Algorithm [76] A different approach to cross-contrast registration is to maximize edge alignment between registered images. While edge-based methods have been used in video and medical image registration, they typically focus on rigid or affine transformations. The EVolution algorithm advances this concept by proposing an edge-based variational method for multi-modal image registration, specifically addressing deformable transformations. The underlying assumption is that physiological structures exhibit sufficient contrast in both image modalities. By prioritizing edge alignment, the EVolution algorithm is particularly effective in scenarios where intensity-based methods may falter, providing a robust solution for cross-contrast registration tasks.

Other Traditional Methods Several traditional methods are widely used as baselines in medical image registration tasks, each offering distinct approaches. SyN (Symmetric Normalization) from the ANTs toolkit [77] optimizes a symmetric cost function for accurate and topology-preserving deformations, making it a robust benchmark. Elastix [78] provides a versatile toolbox for intensity-based registration, supporting rigid, affine, and B-spline transformations, and is known for its flexibility in combining different optimization strategies. NiftyReg [79] leverages GPU acceleration for fast free-form deformation using a cubic B-spline model, making it highly efficient for complex tasks. DEEDS [80] uses discrete optimization with Markov Random Fields (MRFs) to handle complex deformations, particularly excelling in tasks like lung CT registration and ventilation estimation.

Deep Learning-Based Registration Approaches

In recent years, deep learning has emerged as a powerful tool for medical image registration, offering significant improvements in speed and automation compared to traditional methods. These approaches typically leverage convolutional neural networks (CNNs) to predict deformation fields, allowing for efficient and accurate alignment of images, particularly in multi-modal settings.

VoxelMorph [81] VoxelMorph is one of the most prominent deep learning-based frameworks for medical image registration. It uses a CNN to directly predict the deformation field between moving and fixed images. This method significantly reduces registration time during inference, as it eliminates the need for iterative optimization. VoxelMorph has demonstrated strong performance across various tasks, including brain MRI registration, by learning complex deformations from training data. Its architecture is designed to be adaptable, with variants that incorporate both supervised and unsupervised learning paradigms.

DeepReg [82] Another influential deep learning-based framework is DeepReg, which extends the capabilities of VoxelMorph by providing a more flexible and modular design. DeepReg supports both pairwise and groupwise registration tasks, and it can handle multi-modal images, such as MRI and CT scans, by incorporating various loss functions that are tailored for different registration tasks. DeepReg emphasizes the importance of integrating domain-specific knowledge into the learning process, making it particularly effective for complex medical imaging scenarios. It has been applied successfully in various applications, including abdominal and lung imaging.

Other Deep Learning Methods In addition to VoxelMorph and DeepReg, several other deep learning methods have made significant contributions to the field of multi-modal image registration. LapIRN employs a hierarchical approach to registration, using deep networks to progressively refine the deformation field at multiple scales. This method has shown strong performance in tasks like lung and brain imaging, where fine details are crucial. ConvexAdam [83, 84] combines traditional optimization with CNN-based semantic features, allowing for more accurate inter-subject registration by leveraging both deep learning and conventional techniques. It achieved state-of-the-art performance in the Learn2Reg registration challenge [85]. These methods exemplify the growing trend of hybrid approaches that combine the strengths of deep learning with traditional optimization strategies, further enhancing the accuracy and efficiency of medical image registration.

Implemented Registration Pipeline

After resampling the images to a common voxel size, we apply a two-step registration pipeline to achieve accurate alignment of images from different modalities. This pipeline includes rigid and deformable conventional registration techniques. Despite the advancements in deep learning, the Learn2Reg challenge [85] has demonstrated that there is virtually no difference in computational speed between learning-based and optimization-based methods when GPU acceleration is applied. Furthermore, learning-based methods have shown significant improvements primarily in inter-subject registration tasks with strong label supervision. In contrast, for our multi-modal registration task without strong label supervision, conventional methods like EVolution, an edge-based algorithm, are particularly well-suited for aligning MR and CT images. The fine structures of interest in our study, such as vessels and bronchi, benefit from the EVolution algorithm's focus on maximizing edge alignment, making it an effective choice for cross-contrast registration where intensity-based methods may struggle. Moreover, the need for large annotated datasets and the complexity of training deep learning models [86] were also factors in our decision to focus on a more traditional, optimization-based approach.

Rigid Registration Following resampling, we performed a rigid registration to provide a coarse alignment. This initial step involved estimating a rigid transformation using a gradient-driven optimization scheme [87]. The rigid registration aligns images based on translation and rotation, setting the stage for more precise alignment in the next phase.

Deformable Registration For detailed alignment, we applied the EVolution algorithm [76, 88]. This method uses a patch-based approach with a focus on edge alignment and includes a diffusion regularization term to ensure smooth and realistic deformations. To maintain anatomical accuracy and avoid unrealistic distortions, we ensured a diffeomorphic transformation by minimizing the inverse consistency error [89, 75].
Manual verification was finally performed to identify and exclude volumes with significant misalignment due to differences in respiratory phase.

Field of view standardization

After the registration process, the next step is to standardize the field of view (FOV) across both modalities. The FOV often differs between imaging modalities, with MRI scans typically capturing a broader FOV compared to CT scans. In lung MRI, the arms are more likely to appear due to the broader FOV and the positioning of the patient during

the scan. In contrast, lung CT scans generally exclude the arms, focusing specifically on the lungs and positioning the patient in a way that avoids including the arms, as they are not relevant to lung assessment.
To create a uniform dataset and guide the training process, it is essential to standardize the FOV across both modalities, avoiding regions that are present in only one modality. To achieve this, we opted to crop the images around the regions of interest, specifically the lungs. In the literature, various methods are available for effectively segmenting the lung envelope in CT images, including the lungmask method based on a trained U-Net [90].
Given that the CT images have already been resampled and registered to the MRI, we are able to apply the segmentation of the lungs from the CT on the MRI, and apply the same crop to both CT and MR images, ensuring consistency across the modalities. To retain some context outside the lungs while avoiding capturing irrelevant information, we defined a margin of 20 pixels around the segmented lung mask. After this process, volumes were either cropped or zero-padded to have axial slices with a uniform size of 256×256 pixels, depending on the dimensions of the CT lung mask.

2.1.4 Results

After completing all the preprocessing steps, including resampling, registration, field of view standardization, and intensity normalization, we are able to obtain a fully-paired dataset. With this dataset, we could effectively apply supervised GAN-based methods for lung MR to CT synthesis. In addition to enabling paired methods, this dataset also allowed us to evaluate unpaired methods. By comparing the performance of unpaired methods on both an unpaired dataset and the newly paired dataset, we are able to quantify the value of our preprocessing pipeline. In this section, we describe the training process of the models and provide a comparison of their performances in this task using standard image processing metrics. This evaluation helps to quantify the feasibility of synthesizing CT images from MR data and assess the impact of our preprocessing steps.

Training For the unpaired dataset, only the resampling and intensity normalization steps were used as preprocessing, ensuring similar resolution between modalities. For the paired dataset, all preprocessing steps were applied, including resampling, registration, field of view standardization, and intensity normalization. For both datasets, axial slices of size 256×256 were extracted from the MR and CT volumes and saved as 8-bit NumPy arrays. The resulting datasets were then partitioned into a training set comprising 82 patients (33,002 slices) and a testing set of 28 patients (10,725 slices).
All models were trained using the Adam optimizer [91], with a batch size of 1. The generator employed ReLU activation with a learning rate of 0.0001, while the discriminator used LeakyReLU activation and a learning rate of 0.0004, following the guidelines in [92]. The training process included 100 epochs for paired methods and 200 epochs for unpaired methods. Models were trained using available implementations, and training took between 50 and 100 hours on a 12GB NVIDIA Titan X GPU. Inference on a single GPU for MR to CT translation with a trained model took approximately 30 seconds per volume.

Qualitative Evaluation A comparison of synthesized CT images with the corresponding input UTE MR and ground truth CT is displayed in Figure 2.4. All the models tested produce images with properties and resemblance to a real CT scan. However, the

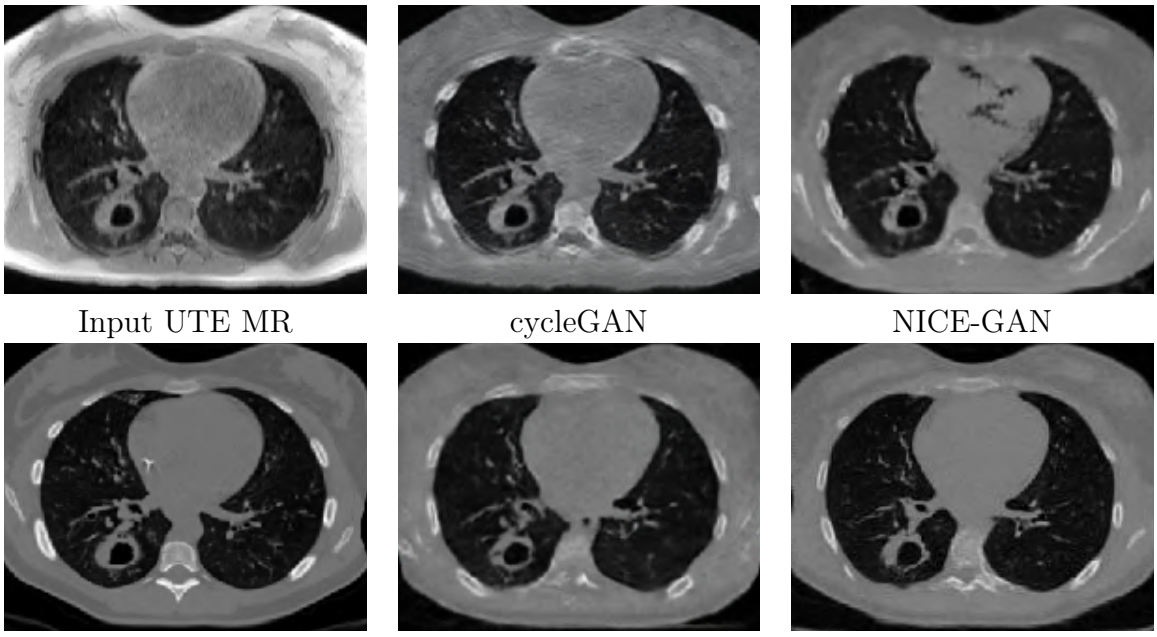


Figure 2.4: Comparison between input UTE MR, ground truth CT and synthetic CT from various models in axial view.

differences between the models are primarily evident in the details of the reconstruction. For the unpaired methods, NICE-GAN exhibits some artifacts, particularly in the regions between the two lungs and near the heart. Additionally, the detail within the lungs is moderate, with some noise present in the image. CycleGAN also shows artifacts, especially in the lower part of the image and around the bones. The reconstruction quality within the lungs is average, with some areas lacking clarity and sharpness.
In contrast, the paired methods generally offer better image quality. pix2pix produces images with good overall quality but lacks detail and finesse within the lungs. Some bronchial structures are exaggerated, appearing wider than those seen in the corresponding MRI and ground truth CT. Despite these issues, the images are relatively clear and free from major artifacts.
pix2pixHD stands out for its high-quality image synthesis. The images generated by this model feature fine detail both within and outside the lungs. The bronchial structures are well-reconstructed, however some vessels appear finer than those on the corresponding MRI and CT. Overall, pix2pixHD produces the most accurate and visually convincing images among the models tested.

Quantitative Evaluation To ensure the fidelity of the reconstructions and validate the initial qualitative assessment, it is essential to evaluate the synthesized images using quantitative metrics. These metrics provide an objective measure of image quality and enable a rigorous comparison of the performance of the different models. We evaluated the synthesized images using standard image processing metrics, including mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM)
To obtain these metrics, 3D CT volumes were reconstructed by stacking the 2D slices. The metrics were computed by comparing the synthesized CT volumes with the ground truth

CT volumes, specifically within the lung masks, to focus on the reconstruction quality of critical lung structures such as vessels and bronchi.
As shown in Table 2.1, the quantitative metrics align with the qualitative observations. Paired methods, such as pix2pix and pix2pixHD, deliver better reconstruction quality compared to unpaired methods. Notably, pix2pixHD outperforms other models, providing the lowest MAE and highest SSIM, indicating superior accuracy in capturing details within the lungs. Furthermore, we can note that unpaired methods also benefit from using the paired dataset, as both CycleGAN and NICE-GAN showed improved performance when trained on the registered dataset compared to the unregistered one. These results underscore the importance of the preprocessing pipeline in enhancing the overall quality of the synthesized images.

Registered dataset	**MAE**	**PSNR**	**SSIM**
pix2pix	81.53 ± 12.73	25.94 ± 1.02	0.786 ± 0.017
pix2pixHD	**75.88 ± 11.62**	**26.56 ± 0.97**	**0.798 ± 0.015**
cycleGAN	91.94 ± 12.46	25.21 ± 1.14	0.771 ± 0.021
NICE-GAN	82.49 ± 10.98	26.18 ± 1.08	0.787 ± 0.018
Unregistered dataset			
cycleGAN	95.78 ± 13.29	25.03 ± 1.17	0.768 ± 0.024
NICE-GAN	86.40 ± 11.41	25.97 ± 1.07	0.775 ± 0.019

Table 2.1: Standard image processing metrics for different GAN architectures.

2.1.5 Discussion

In this initial study, we investigated the feasibility of synthesizing lung CT images from UTE MRI using GAN-based image-to-image translation methods. Our dataset consisted of 110 paired MR and CT images, with careful preprocessing steps, including resampling and registration, to ensure consistency across modalities. The registration step was crucial, as it not only improved the performance of unpaired methods but also enabled the use of paired GAN models, which are known to deliver better results in image synthesis tasks [24]. This underlines the importance of aligning contours and simplifying the translation process for better performance in cross-modality synthesis [48].
We compared unpaired methods like CycleGAN and NICE-GAN to paired methods such as pix2pix and pix2pixHD on both registered and unregistered datasets. As expected, paired methods outperformed unpaired methods in terms of image quality. Among the paired methods, SPADE, which combines pix2pixHD with spatially adaptive normalization, yielded the best results. The novel normalization in the SPADE approach [44] helps preserve content and mitigates the common issue of GANs hallucinating anatomical features, which can lead to false positives.
Our findings demonstrate the feasibility of MR to CT lung image synthesis, with promising results in terms of image quality. However, caution is necessary regarding the potential for GAN-generated images to introduce false positives or hallucinated features [53], particularly in clinical contexts. Although our methods show strong performance, they do not inherently control for the generation of such artifacts, which is a critical consideration in medical imaging. As a result, there is a need for further research and development to refine these models and implement mechanisms to ensure that the synthesized images are

accurate and reliable for clinical use. In conclusion, this study confirms that paired image-to-image translation GANs, particularly pix2pixHD, can effectively synthesize lung CT images from UTE MRI. The choice of paired methods aligns with previous research, indicating that unpaired methods can introduce artifacts and distort fine structures. While the results are encouraging, additional work is needed to refine these methods and ensure their robustness in various clinical scenarios, with a focus on minimizing hallucinations and false positives in the generated images.

2.2 Optimizing Preprocessing for Improved Synthesis

As a first step toward optimizing the translation of MR to CT images, we focus on the preprocessing pipeline. In medical imaging, where accurate anatomical representation is essential, the choice of preprocessing parameters can significantly impact the quality of the synthesized images. This section details the preprocessing strategies employed in our study, including the evaluation of a conditional registration step, different resampling techniques, normalization approaches, and the effects of intensity quantization on the final output. By refining these steps, we aim to enhance the overall performance of GAN-based synthesis and reduce potential artifacts.

2.2.1 Enhancing Registration using Lung Masks

In this section, we will detail a new approach for registration that incorporates a constraint based on the lung envelope from both modalities. By focusing on the lung regions, we aim to enhance the accuracy of the registration process, ensuring that the structures of interest, such as vessels and bronchi, are correctly aligned while minimizing the influence of less relevant outer regions.

Segmentation of the lungs in MR and CT

Our primary objective in lung MR to CT translation is to accurately generate structures within the lungs, specifically the lung parenchyma. After discussions with radiologists, it became clear that the outer regions, such as the bones and mediastinum, hold less importance for the analysis of anatomical structures such as vessels and bronchi, particularly in the context of conditions like cystic fibrosis. These outer areas can also pose challenges during registration, as they are influenced by respiratory movement, leading to potential misalignment. To address these issues and focus on the structures of interest within the lungs, we opted to apply a conditioned registration approach, constraining the registration process to lung regions only.
For the CT modality, obtaining a precise lung mask is straightforward, thanks to state-of-the-art algorithms such as the pretrained U-Net lungmask model [90]. This allowed us to quickly generate accurate lung masks for CT images, which could then be used to guide the registration process.
However, UTE MRI lung segmentation is more challenging due to the lower signal intensity and contrast in lung regions, and currently, there are no established methods in the state of the art for this specific task. This is likely due to UTE MRI being a relatively recent imaging technique, and very few datasets are available for lung segmentation in this modality [68]. To address this gap, we developed a specialized pipeline for MRI lung mask generation, leveraging masks from registered CT, expert radiologist corrections and

the nnU-Net framework.
The nnU-Net [93] is a state-of-the-art segmentation method, and the first framework explicitly designed to handle the dataset diversity found in biomedical imaging. nnU-Net automates key decisions needed to design a robust segmentation pipeline, making it an ideal candidate for tackling challenging tasks like lung MRI segmentation. As a standardized benchmark in the field, it provides a plug-and-play solution that can be applied across various datasets without manual intervention, streamlining the development of segmentation methods. nnU-Net's modular structure allows for easy integration of new architectures, enabling rapid evaluation of modifications across different datasets in a consistent environment. Our pipeline, as depicted in Figure 2.5, involves two iterations of nnU-Net models, each designed to progressively enhance the segmentation quality.

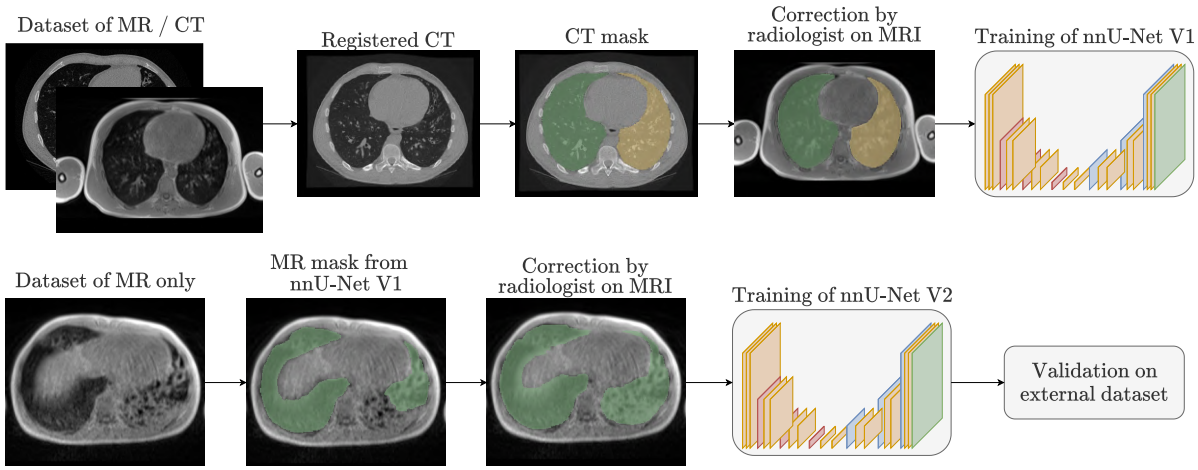


Figure 2.5: Illustration of our MR segmentation pipeline. The CT mask is obtained using the lungmask model [90], and propagated on the paired MRI to ease the manual annotations from the radiologist.

First Iteration: Creation of an Annotated Database (nnU-Net V1) The initial step involved creating an annotated dataset for segmenting lung masks on MRI. We utilized our database of 110 patients, all of whom had MRI and CT scans performed on the same day, primarily for cystic fibrosis. After registering the CT images to the corresponding MRIs, we used lung segmentations obtained from the CT scans using the lungmask algorithm as a preliminary draft for the MRI lung masks. This draft served as a starting point for the radiologist, who then corrected and refined these masks to align with the MRI anatomy. This semi-automated approach significantly reduced the manual effort required from the radiologist, making the process more efficient. The corrected masks from this step were then used to train the first neural network model, referred to as nnU-Net V1.

Second Iteration: Expansion of the Dataset (nnU-Net V2) In the second iteration, we expanded our dataset by adding new patients who had only undergone MRI scans without corresponding CT scans. These patients were diagnosed with either cystic fibrosis or interstitial lung disease, providing a broader range of pathological variations. We employed the nnU-Net V1 model to perform an initial segmentation of the lungs in these new MRI scans. The rough segmentations generated by V1 were then refined by

an expert radiologist through manual corrections. This iterative process helped create a more comprehensive and diverse dataset, referred to as the V2 dataset, which was used to train the second neural network model, nnU-Net V2. The nnU-Net V2 model outperformed its predecessor, largely due to the increased size of the training dataset and the inclusion of a wider variety of lung diseases, which enhanced the model's robustness and generalization capabilities.
Figure 2.6 illustrates the differences in segmentation results between the V1 and V2 models on a patient with lung alterations. The V2 model successfully segmented the lungs where the V1 model encountered difficulties to capture the alterations, highlighting the improvements achieved through the iterative refinement process.

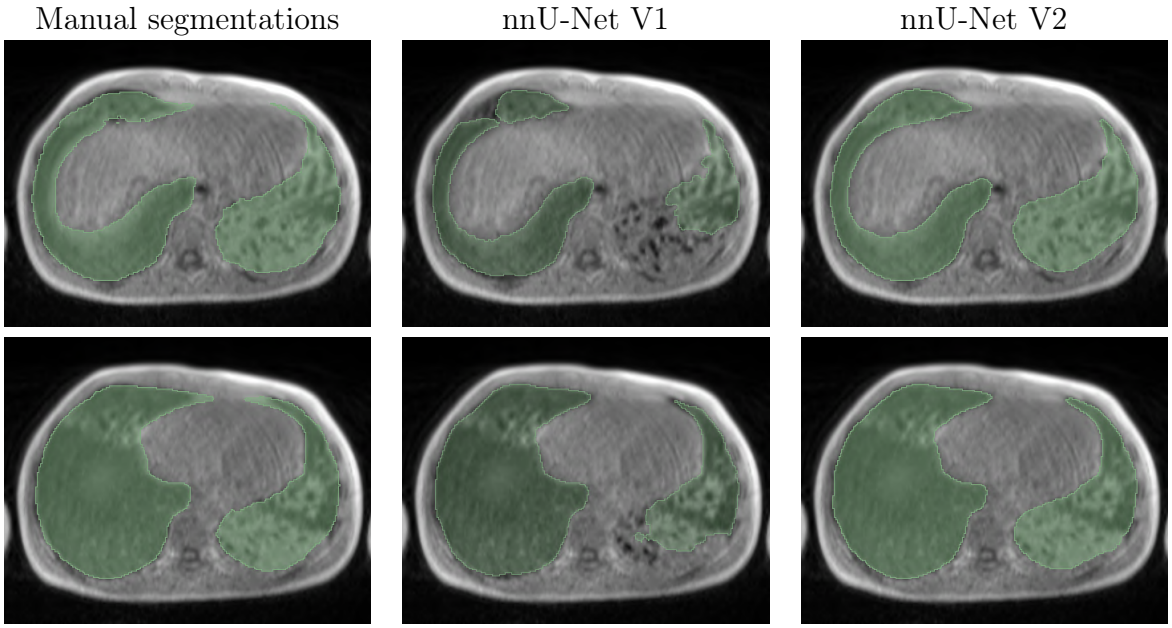


Figure 2.6: Comparison between ground truth MRI lung mask and outputs from models V1 and V2. Axial slices of the same patient with significant alterations of the left lung are shown.

The key advantage of our pipeline is its ability to provide an accurate pre-segmentation of the lungs for each image, either based on the recalibrated CT segmentation or a coarse segmentation from a trained model. This approach offers several benefits: it significantly reduces the manual annotation workload for radiologists, which can be particularly tedious, and it allows for more accurate final segmentations. The coarse segmentation is performed in 3D and relies on image contours, enabling the radiologist to focus on refining the segmentation rather than starting from scratch. This method also mitigates the risk of staircase-like artifacts, which are common when segmentations are manually drawn in 2D.
As part of our collaboration with the Bordeaux Hospital Center, the evaluation of our deep learning model for UTE MR lung segmentation is currently undergoing clinical assessment, further validating the effectiveness and applicability of our approach in a real-world clinical setting.

The nnU-Net V2 can now be integrated into our registration pipeline, serving as an input for our new conditional registration process, which utilizes the generated lung masks from both MRI and CT modalities to enhance the accuracy and effectiveness of the alignment.

Conditional Registration

At the beginning of this chapter, we described our multi-modal registration pipeline, which leverages conventional methods like the EVolution algorithm for aligning MR and CT images. This algorithm's edge-based approach is particularly well-suited for our task of aligning fine structures such as vessels and bronchi, as it prioritizes edge alignment—a critical factor in lung imaging. Our registration process begins with rigid registration to achieve coarse alignment, followed by deformable registration using EVolution to ensure a detailed match. However, to further enhance the accuracy of this alignment, we introduce a conditional registration approach that focuses on the lung regions specifically.
The principle behind this conditional registration is to focus on accurately aligning the structures inside the lungs, such as vessels and bronchi. By concentrating on these areas, the edge-based EVolution method becomes particularly appropriate for guiding the registration.
Typically, a registration method relies on two main components: a data attachment term and a regularization term [94]. The regularization term ensures smoothness and avoids unrealistic deformations and does not require modification in this approach. However, we modify the data attachment term to constrain the registration process to voxels within the lungs, using the lung masks generated by nnU-Net for MRI and lungmask for CT. This adjustment ensures that the registration process prioritizes aligning structures within the lungs, minimizing the influence of irrelevant outer regions.

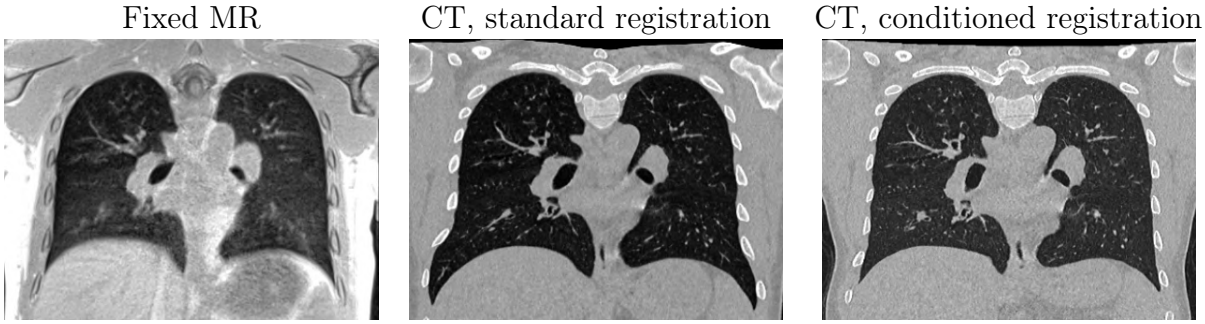


Figure 2.7: Comparison of registration methods: Standard registration struggles with lung alignment due to size differences between modalities, while conditioned registration improves alignment by focusing on internal lung regions.

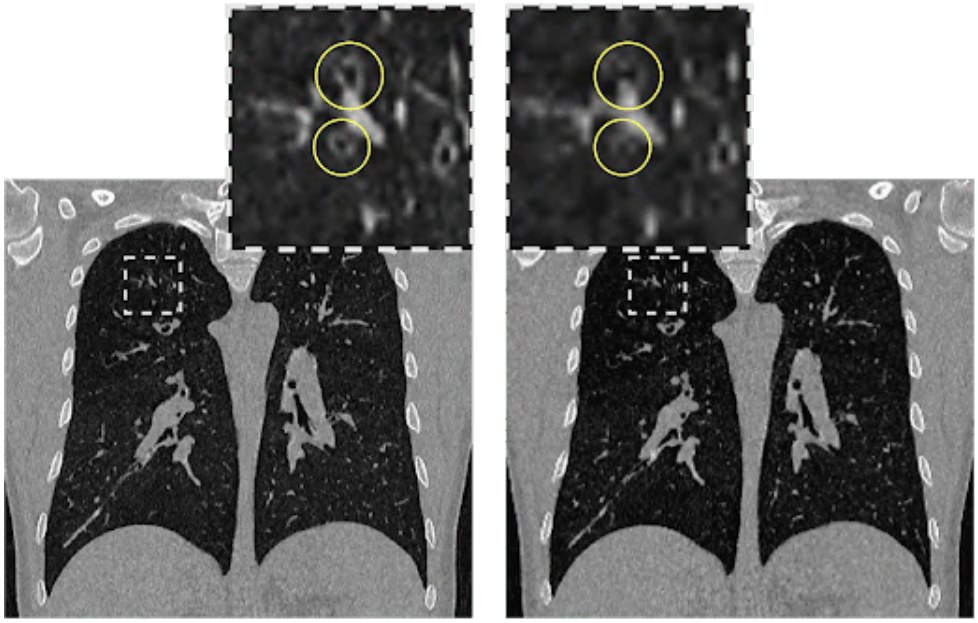
Figure 2.7 presents a comparison between the standard global registration approach and our conditional registration method. In the case of a CT acquired during the end-breathing phase (i.e., with lungs inflated with air), the global method struggles to properly align the lungs, particularly the inferior lobe of the right lung, due to the significant differences in lung sizes between modalities. The conditional registration method, however, corrects this issue by focusing on the lung interiors, allowing the registration to guide itself based on the regions that are supposed to match.

2.2.2 Preprocessing Strategies

Resampling to CT voxel size

In multimodal registration, the conventional approach is to resample images based on the voxel size of the fixed image, which typically has the highest resolution. This method

ensures that the registration process benefits from the maximum level of detail and accuracy. However, in our case, this standard approach presents a significant challenge: the MRI, which serves as the fixed image, has a voxel size of $1 \times 1 \times 1$ mm^3, while the CT volume, the moving image, has a higher resolution with a voxel size of $0.6 \times 0.6 \times 0.6$ mm^3. Resampling the CT to match the MRI resolution would result in a loss of crucial information, particularly fine anatomical details, during the interpolation process, as shown in Figure 2.8



CT $0.6 \times 0.6 \times 0.6$ mm^3 CT $1 \times 1 \times 1$ mm^3

Figure 2.8: Visual comparison of identical CT slices at different voxel spacing, with zoomed regions that highlight bronchi, circled in yellow. Pulmonary bronchi are nearly indistinguishable in the $1 \times 1 \times 1$ mm^3 version due to the lower resolution of the image.

To address this issue, we propose a deviation from the state-of-the-art approach by defining a common grid based on the higher CT resolution of $0.6 \times 0.6 \times 0.6$ mm^3. This strategy of preserving the original resolution of the CT helps to avoid the loss of important details that are critical for accurate image translation. By maintaining the CT images at their highest quality, which serve as the ground truth for the translation, we ensure that the GAN models have high-quality reference data, potentially leading to better translation results.
Upsampling the MRI from $1 \times 1 \times 1$ mm^3 to $0.6 \times 0.6 \times 0.6$ mm^3 may be viewed critically because it introduces interpolated data. However, in this case, the interpolation does not create new intensities but rather estimates them based on the existing data, ensuring consistency. Importantly, this upsampling process is reversible, allowing us to return the MRI data to its original resolution without losing any information. In contrast, downsampling the CT results in a permanent loss of data.

Intensity Normalization

CT and MR modalities have fundamental differences that must be taken into account when normalizing intensity values. CT intensity values are defined in Hounsfield units (HU) and have a physical meaning, whereas MR intensity values strongly depend on

acquisition parameters. Therefore, methods used for intensity normalization must be tailored to the specific characteristics of each modality.
In our study, CT intensities are cropped to [-1000; 2000] HU window to remove irrelevant values from the table or background and rescaled to [-1; 1] using the same window limits. MR intensity inhomogeneities, also known as bias field, are first corrected using the popular N4 bias field correction algorithm [95]. MR values are then normalized using z-score, i.e. zero mean and unit variance, cropped to $[-3\sigma\ ;\ 3\sigma]$ to remove outliers, σ being the standard deviation, and rescaled to [-1; 1] based on minimum and maximum intensities. Nyul histogram matching [96] was also considered, but the findings of Reinhold et al. [71], along with our preliminary results, indicated that the synthesis process was robust to the choice of MR normalization method used. As a result, we opted for a traditional Z-score normalization approach with outliers removal.

Impact of intensity quantization

In this study, we also investigate the impact of the bit depth of input medical images on the performance of a GAN in lung MR to CT translation. Medical images are typically encoded in 16 bits, but in practice, they often contain only 12 bits of significant data [97]. For example, in CT scans, the Hounsfield unit (HU) values, which represent tissue density, are mapped onto 4096 possible values, ranging from -1024 HU to +3071 HU. However, many GAN implementations in the field are derived from natural image datasets, which are commonly encoded in 8 bits. To explore the implications of this, we created two datasets: one using the standard 8-bit images, as is common in current state-of-the-art implementations, and another using 16-bit images to preserve the full range of medical data. This comparison allows us to evaluate whether the higher bit-depth images can improve the performance of the GAN for thoracic CT synthesis, potentially capturing more detailed intensity variations that are crucial for accurate image translation.

2.2.3 Results

In previous sections 2.2.2 and 2.2.1, we described various preprocessing optimizations and a new conditional registration method. In this section, we will compare and evaluate these different optimizations through a series of experiments. We will elaborate on the datasets configurations, training models, and parameters, along with a comparison between the best-performing model from the section, pix2pixHD, and its variant using SPADE adaptive normalization. Qualitative and quantitative results will then be presented, focusing on the reconstruction of clinically significant structures such as bronchi.

Datasets Configurations

In this study, several key changes were implemented in the experimental settings compared to the previous section 2.1. Firstly, we updated the intensity normalization process for MRI, ensuring better standardization across varying acquisition parameters. Additionally, we enhanced the registration process by incorporating conditional registration based on input masks. For UTE MRI, the lung masks were generated using our trained nnU-Net [93], while for CT, the lung masks were derived from the available lungmask algorithm.
To facilitate a comprehensive comparison, we created three distinct datasets, as described in Figure 2.9. The first dataset maintained a voxel size of $1 \times 1 \times 1$ mm^3, consistent with our previous experiments, and included axial slices resized to 256 x 256 pixels. The

second dataset adopted a higher resolution with a voxel size of $0.6 \times 0.6 \times 0.6$ mm^3, producing axial slices resized to 512 x 512 pixels. Within this higher resolution dataset, we generated two versions: one with 8-bit images and another with 16-bit images. All datasets utilize the new conditional registration approach, employing MR and CT lung masks for enhanced alignment.

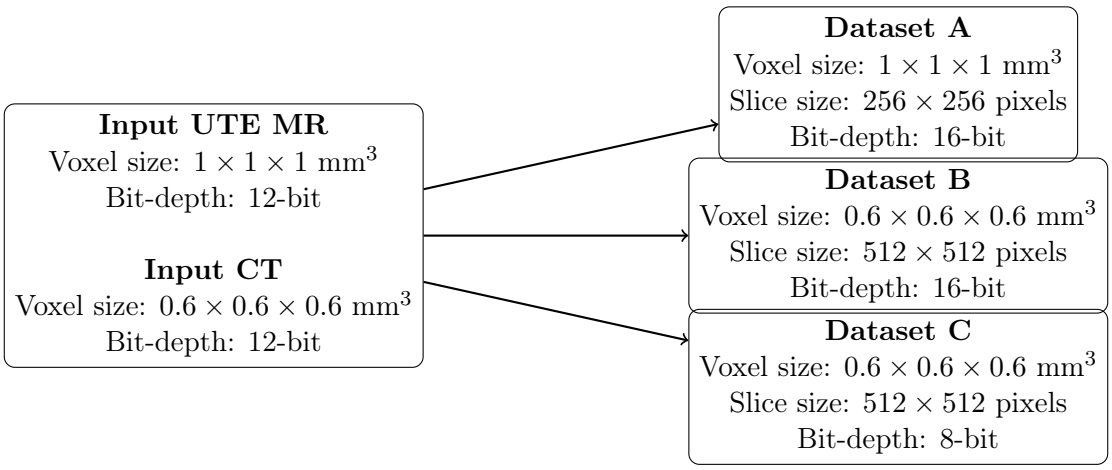


Figure 2.9: Overview of input properties and resulting datasets

Training We kept the same training parameters as in section 2.1, focusing exclusively on the top-performing model, namely pix2pixHD. Additionally, we experimented with enhancing pix2pixHD using Spatially-Adaptive Normalization (SPADE).
SPADE aims to replace traditional batch normalization, which normalizes activation outputs based on a unit Gaussian distribution, with a Conditional Batch Norm. This is achieved by projecting the input image onto an embedding space and then convolving it to produce modulation parameters. The idea behind using SPADE in our MR to CT translation task is to mitigate the typical hallucination tendencies of GANs. By incorporating conditional normalization based on the input image, we aim to prevent the model from deviating too far from the input and ensure that generated features remain closely aligned with the original image.
Both pix2pixHD and pix2pixHD with SPADE normalisation were trained under the same conditions, ensuring consistency in the experimental setup while allowing us to assess the impact of the updated preprocessing strategies and dataset configurations.

Qualitative Evaluation

Figure 2.10 visually compares an axial slice of the input MR, the ground truth CT, and the synthetic CT generated by pix2pixHD alone and with the SPADE normalisation, using different voxel sizes. The results obtained with 0.6 mm isotropic voxels exhibit enhanced contrast and sharpness, allowing for a more accurate distinction of vessels and bronchi within the lung parenchyma compared to the 1mm voxel size.
The pix2pixHD 0.6mm output presents a noise-free image, an accurate reconstruction of the bones and body, but seems to deviate from the CT at the level of the lung parenchyma, in particular with the appearance of bronchi not visible on the ground-truth CT. Conversely, the SPADE 0.6mm output presents a slightly less sharp image, a less accurate

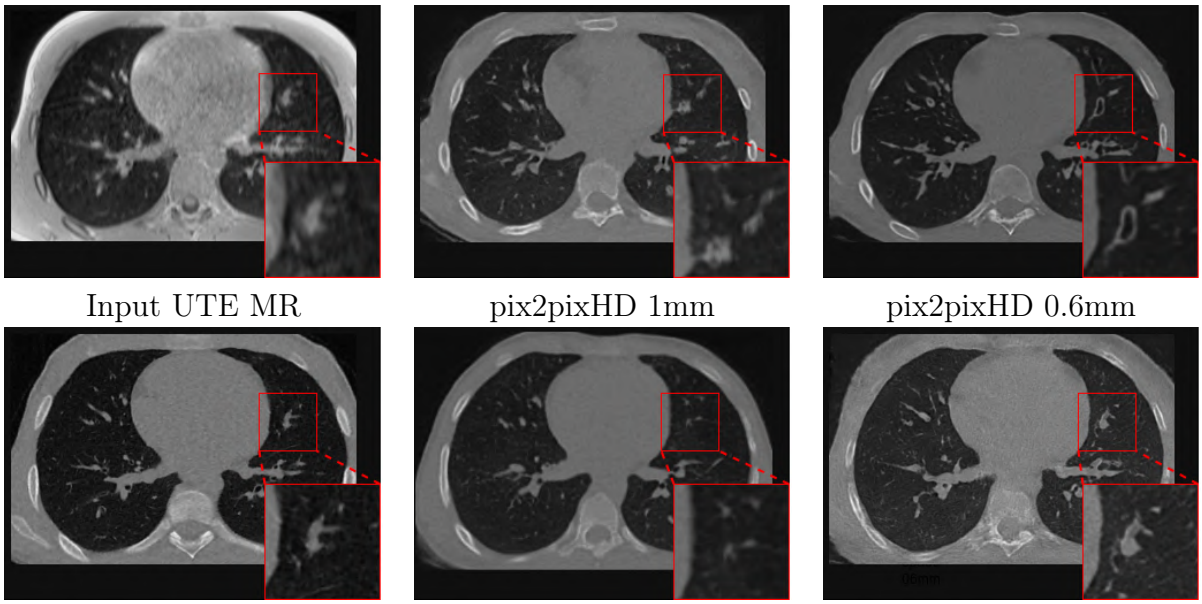


Figure 2.10: Comparison between MR axial slices, ground truth CT and synthetic CT from pix2pixHD and SPADE models using the $1 \times 1 \times 1$ mm^3 and $0.6 \times 0.6 \times 0.6$ mm^3 datasets, with zoomed-in views of bronchi reconstruction. The pix2pixHD 0.6mm model demonstrates some false positives, with the appearance of bronchi not present in the ground-truth CT. In contrast, the SPADE 0.6mm model provides a more accurate reconstruction of fine structures.

reconstruction of the bones, but a very faithful synthesis of the lung parenchyma, in particular at the level of the vessels and bronchi.

Quantitative Evaluation

Quantitative evaluation was performed using mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM). These metrics were computed between synthesized CT images and ground truth CT images, focusing on the region within the intersection of the lung masks to isolate the lung parenchyma for analysis.
Results, presented in Table 2.2, are in line with visual findings. pix2pixHD with SPADE normalization layers consistently achieves better performance in all image processing metrics compared to pix2pixHD alone, regardless of the dataset. The 0.6mm dataset also demonstrates superior performance to the 1mm dataset. Interestingly, we observe that the dataset with 0.6mm resolution and 8-bit depth delivers results very close to those obtained with 16-bit depth, suggesting that quantization does not really impact the results in our dataset.
Table 2.3 lists a comparison of computational resources required by the networks, including memory cost on the GPU, training time per iteration, and disk storage impact. The table highlights how different voxel sizes and bit depths influence these computational aspects, showing that while using a 1mm voxel size with a 256x256 image resolution results in faster training and lower memory cost compared to a 0.6mm voxel size, the bit depth (8 bits vs. 16 bits) primarily affects the disk storage requirements, with no significant impact on memory footprint or calculation time.

Table 2.2: Mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index (SSIM) between synthesized CT and real CT inside the lungs.

Dataset A (1mm, 16-bit)	**MAE**	**PSNR**	**SSIM**
pix2pixHD	74.14 ± 11.85	26.79 ± 0.94	0.801 ± 0.015
SPADE	71.36 ± 9.79	27.41 ± 0.91	0.809 ± 0.015
Dataset B (0.6mm, 16-bit)			
pix2pixHD	71.51 ± 11.75	27.30 ± 0.98	0.805 ± 0.016
SPADE	**67.53 ± 10.31**	**27.95 ± 0.85**	0.813 ± 0.015
Dataset C (0.6mm, 8-bit)			
pix2pixHD	71.89 ± 11.79	27.24 ± 0.99	0.804 ± 0.015
SPADE	67.76 ± 10.70	27.87 ± 0.88	**0.814 ± 0.014**

Table 2.3: Comparison of networks computational resources in terms of memory cost on GPU and training time per iteration.

Model	**Size of Input Data**	**Memory Cost**	**Time per iter.**
SPADE, 1mm, 16-bit	129KB	3 GB	0.135 s
SPADE, 0.6mm, 8-bit	257KB	4.5 GB	0.316 s
SPADE, 0.6mm, 16-bit	513KB	4.5 GB	0.317 s

2.2.4 Discussion

This study focused on evaluating the impact of different preprocessing strategies on the performance of GAN-based MR to CT translation, specifically targeting lung structures. We explored various preprocessing parameters, including resampling and registration. In terms of resampling, we introduced a new strategy that deviates from traditional approaches by using a 0.6 mm voxel size instead of the standard 1 mm. This approach challenges conventional practices by prioritizing the preservation of fine anatomical details, particularly within the lung parenchyma, which are crucial for accurate image synthesis. Additionally, we enhanced the registration process by incorporating a condition based on lung masks, ensuring that the registration focused on the lung region. For MRI lung segmentation, we developed an innovative pipeline using the nnU-Net framework. This pipeline relied on registered CT lung masks and employed a two-step approach to guide and reduce the radiologist's annotation workload, ultimately improving the accuracy of MRI lung segmentation.

The results from the pix2pixHD and SPADE models across the three datasets confirmed the significance of accurate resampling. The use of a 1x1x1 mm voxel size led to inferior results, supporting our initial assumption that higher-resolution resampling (0.6 mm) is critical for maintaining GAN performance, particularly in preserving fine structures like vessels and bronchi.

Interestingly, the results from the 8-bit dataset were similar to those from the 16-bit dataset, which was somewhat surprising. This suggests that the highly contrasted information in our dataset, such as vessels and bronchi, may not be significantly impacted by the reduced bit-depth. The representation in 8-bit, rather than 16-bit, may have minimal effect on the reconstruction quality.

Finally, the results obtained with the SPADE model consistently outperformed those of pix2pixHD alone, both qualitatively and quantitatively. The combination of pix2pixHD with the spatially adaptive normalization systematically led to better performance, re-

ducing false positives and enhancing the quality of MR to CT translations. This validates the advantage of using SPADE for such tasks, as it appears to significantly improve the model's ability to preserve anatomical accuracy and reduce errors in synthetic CT images. Despite these promising results, it is important to recognize some limitations in our evaluation. The quantitative assessment relied solely on standard image processing with MAE, PSNR, and SSIM. While these metrics provide valuable insights, they may not fully capture the extent of anatomical preservation achieved by each method. Our qualitative evaluation revealed the presence of false positives in certain methods, such as the unpaired methods discussed in the section 2.1 and even in pix2pixHD in this study. This underscores the need for more clinically relevant metrics to evaluate synthetic CT images effectively. Standard quantitative metrics alone do not adequately reflect the clinical accuracy required for these images to be used in practice, highlighting the importance of integrating clinical validation into future evaluations.

2.3 Clinical Assessment

Building on the feasibility study of MR to CT synthesis using paired and unpaired GANs, as well as the subsequent optimization of preprocessing strategies, this section explores the clinical relevance of the synthetic CT images generated by our pipeline. The goal of this assessment is to evaluate whether these synthetic CTs meet the standards necessary for accurate diagnosis and clinical decision-making, particularly in comparison to real UTE MR and CT images.
To accomplish this, we conduct a comprehensive clinical evaluation that not only considers technical image quality metrics but also examines the synthetic CTs' ability to accurately represent clinically significant pulmonary structures. This assessment is crucial for determining the practical utility of synthetic CT images in a clinical setting, especially for conditions like cystic fibrosis (CF).
The clinical assessment is structured as follows: first, we describe the methodology, including the dataset used, the methods for generating synthetic CTs, and the criteria used for evaluation. This is followed by a detailed presentation of the results, highlighting the strengths and potential limitations observed during the assessment. Finally, the discussion section will address the implications of these findings for the clinical use of synthetic CT images, identifying areas for improvement and suggesting directions for future research.

2.3.1 Methodology

In this section, we detail the approach used to generate synthetic CT images using GANs and the methods employed to assess their clinical relevance. The process begins with a description of the dataset, which includes both internal and external data sources to ensure a robust evaluation. Following this, we describe the generation of synthetic CT images using a specialized GAN architecture tailored for this study.
The clinical relevance of these synthetic CTs is then evaluated through a combination of quantitative and qualitative methods. The image quality is assessed by experienced radiologists, who measure key metrics such as contrast-to-noise ratio (CNR) and signal-to-noise ratio (SNR). Additionally, the evaluation includes an analysis of the synthetic CTs' ability to accurately depict structural alterations associated with cystic fibrosis, using well-established scoring systems. This comprehensive approach allows us to rigorously evaluate the potential of synthetic CTs for clinical application.

Dataset Description

For this study, we utilized the same cohort of 110 patients from the previous experiments, with the dataset split into 82 patients for training and 28 patients for testing. In addition to this, an external dataset comprising 46 patients from 5 different institutions was introduced for further evaluation. This external dataset includes paired UTE MRI and CT images, with all imaging performed on the same day to ensure consistency. This external dataset allowed us to assess the generalizability and clinical applicability of the synthetic CT data across different institutions and imaging conditions.

Generation of Synthetic CT using GAN

Synthetic CT images were generated using a 2D GAN architecture, specifically the pix2pixHD model enhanced with Spatially-Adaptive Normalization (SPADE). This setup was chosen based on the best-performing configurations identified in our previous studies. The loss function was defined as a combination of GAN loss, feature-matching loss, and L1 loss.
The preprocessing settings included resampling the images to common voxel size of $0.6 \times 0.6 \times 0.6$ mm^3, ensuring high-resolution input with 16-bit arrays of size 512x512 pixels. Additionally, the registration process was refined by applying a condition based on lung masks, focusing the alignment and synthesis specifically on the lung regions. This combination of advanced GAN architecture and optimized preprocessing was used to generate the synthetic CT images that were subsequently evaluated in the clinical assessment.

Image Quality Evaluation

For quality assessment, two readers with 4 years and 10 years of experience in chest imaging conducted the evaluations according to established methods. Quantitative assessment included measurements of apparent contrast-to-noise ratio (CNR), signal-to-noise ratio (SNR), and noise levels within regions of interest (ROI). These ROIs were manually defined by radiologists within the trachea, main bronchi, and main pulmonary arteries.
Qualitative analysis involved evaluating several visual aspects of the images. This included a total visibility score for central and peripheral airways and vessels (range 0-30), a total sharpness score for vessels and airways (range 0-24), and a total artifact score (range 0-45), which considered five types of artifacts: beam hardening, motion, aliasing, shadowing, and streaks.

Evaluation of Cystic Fibrosis Structural Alterations

Structural alterations such as bronchiectasis, wall thickening, and mucus plugs were assessed on a segmental level. For each lung segment, the presence or absence of these alterations was recorded, allowing for a detailed comparison between real CT, synthetic CT, and UTE MRI. These structural alterations were then further evaluated and quantified using the Bhalla scoring system.
The Bhalla score is a standardized method specifically designed to assess and quantify the severity of lung abnormalities, which is particularly useful in patients with conditions like Cystic Fibrosis. This scoring system evaluates key structural features, including bronchiectasis and peribronchial thickening, by assigning severity scores based on defined criteria. For example, bronchiectasis is scored based on the luminal diameter of

the bronchial tubes compared to adjacent vessels, with higher scores indicating more severe abnormalities. Similarly, peribronchial thickening is assessed by comparing the wall thickness of the bronchial tubes to the adjacent vessels. Additionally, the Bhalla score accounts for the extent of bronchial involvement, noting how far bronchiectasis or mucus impaction extends within the bronchial tree. The total Bhalla score starts at 25, with points subtracted based on the severity of observed abnormalities, making it a comprehensive tool for evaluating lung disease severity.
To ensure the reliability of the scoring, evaluations were performed independently by multiple readers and repeated to assess both intra-observer (within the same observer) and inter-observer (between different observers) reproducibility. This approach ensures that the Bhalla scoring system provides consistent and accurate results across different assessments, making it a valuable tool for tracking disease progression and guiding treatment decisions in clinical practice.

2.3.2 Results

Building on the methodology outlined in section 2.2, where we described the generation of synthetic CT images using optimized GAN-based techniques and the rigorous evaluation criteria applied by expert radiologists, we now present the findings of our clinical assessment. This section details both the quantitative and qualitative outcomes of the image quality evaluation, as well as the correlations of structural alterations in cystic fibrosis patients, comparing synthetic CT images with real CT and UTE MRI. The results provide insight into the effectiveness of our approach in producing clinically viable synthetic CT.

2.3.3 Image Quality

In the test cohort, the signal-to-noise ratio (SNR) was similar across all imaging modalities, indicating no significant differences, as shown in Table 2.4. However, the contrast-to-noise ratio (CNR) was markedly higher in synthetic CT images compared to UTE MRI. Additionally, real CT scans exhibited a higher overall noise level than MRI scans and synthesized CT. This can be due to the use of low-dose CT protocols aimed at reducing radiation exposure, which typically increases image noise.
For the qualitative assessment of image quality, UTE MRI demonstrated lower scores in contour sharpness, indicating more blurring, when compared to both real CT and synthetic CT. Synthetic CT also achieved the lowest total artifact score among the modalities evaluated. Notably, beam hardening artifacts were observed only in CT scans for a significant portion of patients, while slight-to-moderate motion artifacts were present in some UTE MR images due to imperfect respiratory synchronization.

2.3.4 Correlations of Structural Alterations

In the test dataset, structural alterations were evaluated in 504 lung segments from 28 cystic fibrosis (CF) patients. Results are presented in Table 2.5. The comparison revealed that synthetic CT had a better agreement with real CT than UTE MRI in assessing key structural changes such as bronchiectasis, wall thickening, and mucus plugs. This superior performance was further highlighted by a reduction in both false negatives and false positives. For example, UTE MRI missed a higher percentage of lung segments with

Table 2.4: Image quality evaluation for test and external datasets using quantitative and qualitative assessments. Values are presented as median [interquartile range].

Test Dataset (n=28 patients)	**UTE MRI**	**Synthetic CT**	**CT**
Quantitative analysis			
Contrast-to-noise ratio	9.3 [6.6, 35]	303 [221, 382]	247 [210, 323]
Signal-to-noise ratio	88 [87, 91]	88 [84, 92]	88 [86, 91]
Overall Noise	22 [18, 31]	26 [22, 30]	42 [32, 50]
Qualitative analysis			
Total visibility score	27 [24, 29]	29 [27, 30]	29 [27, 30]
Total sharpness score	10.5 [9, 12]	22 [18, 24]	22 [18, 24]
Total artifact score	0.5 [0, 1]	0 [0, 0]	1 [0, 1]
External Dataset (n=46 patients)	**UTE MRI**	**Synthetic CT**	**CT**
Quantitative analysis			
Contrast-to-noise ratio	13 [7, 20]	364 [258, 448]	245 [155, 324]
Signal-to-noise ratio	90 [88, 93]	90 [88, 92]	90 [87, 92]
Overall Noise	21.5 [14, 45]	23 [18, 28]	39 [20, 46]
Qualitative analysis			
Total visibility score	28 [26, 29]	29 [28, 30]	29 [28, 30]
Total sharpness score	11 [9, 18]	22 [20, 24]	22 [18, 24]
Total artifact score	1 [0, 3]	0 [0, 0]	0.5 [0, 1]

Table 2.5: Bhalla score

Test dataset (n=28)	**Ref CT vs. UTE MRI**	**Ref CT vs. Synthetic CT**	**Ref CT vs. CT**
Junior reader			
ICC	0.70	0.92	0.93
95% CI	0.45, 0.84	0.83, 0.96	0.85, 0.97
Senior reader			
ICC	0.96	0.99	N/A
95% CI	0.93, 0.98	0.99, 0.99	N/A
External dataset (n=46)	**Ref CT vs. UTE MRI**	**Ref CT vs. Synthetic CT**	**Ref CT vs. CT**
Junior reader			
ICC	0.77	0.93	0.93
95% CI	0.60, 0.87	0.88, 0.96	0.88, 0.96
Senior reader			
ICC	0.95	0.97	N/A
95% CI	0.92, 0.97	0.94, 0.98	N/A

bronchiectasis and more frequently falsely identified mucus plugs compared to synthetic CT. In the external dataset, which included 828 lung segments from 46 CF patients, synthetic CT again outperformed UTE MRI, especially in aligning with real CT findings for bronchiectasis. The results confirmed that synthetic CT reduced inaccuracies in depicting structural alterations, encompassing both severe and mild cases.

The Bhalla score analysis supported the superior performance of synthetic CT. For the junior reader in the test dataset, UTE MRI Bhalla scores showed substantial concordance with the real CT reference scores (ICC = 0.70), whereas synthetic CT demonstrated excellent concordance, comparable to real CT (ICC = 0.93). In the external dataset, UTE MRI continued to show substantial concordance, while both synthetic CT and real CT exhibited almost perfect agreement. The senior reader achieved excellent concordance for both UTE MRI and synthetic CT Bhalla scores with the real CT reference in both datasets.
Finally, intra-observer reproducibility was nearly perfect for both junior and senior readers across all imaging modalities, underscoring the consistency and reliability of the assessments.

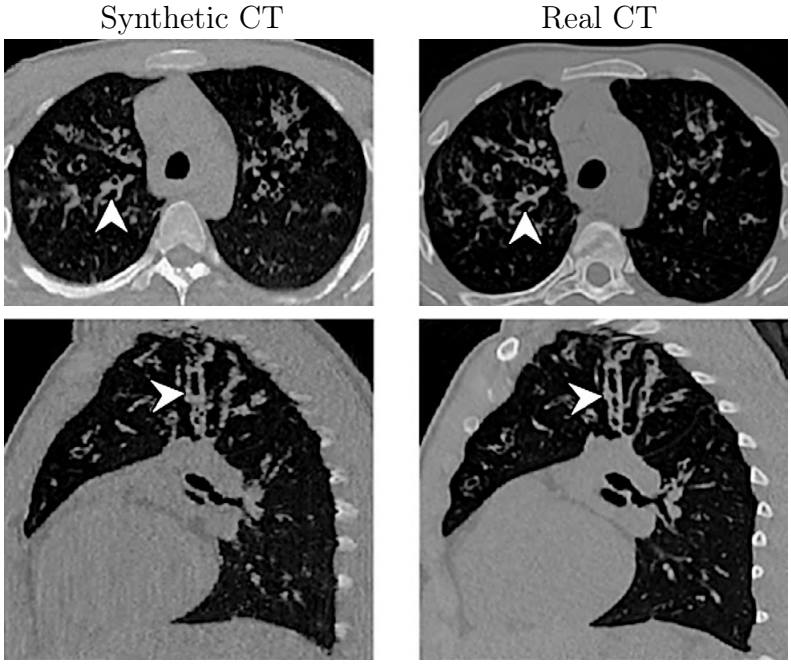


Figure 2.11: Comparison between UTE MRI, synthetic CT, and real CT on a patient with cystic fibrosis from the external dataset. Axial (top) and sagittal (bottom) slices are presented. White arrowheads depict bronchiectasis detected by the radiologist, which were closely reconstructed in the synthetic CT compared to the real CT. A staircase effect is visible in the sagittal slice, caused by the use of 2D slice-wise reconstruction methods.

2.3.5 Discussion

This section aimed to assess whether synthetic CT images generated from UTE MRI could meet the rigorous standards required for accurate diagnosis and clinical decision-making, compared to traditional UTE MRI and CT images. To achieve this, we defined both qualitative and quantitative evaluation criteria, focusing on image quality assessment and the evaluation of cystic fibrosis (CF) structural alterations and severity.
The results of our study demonstrated that synthetic CT images produced higher contrast-to-noise ratios, fewer artifacts, and overall better image quality than UTE MRI. The synthetic CT images showed a significant reduction in common artifacts such as beam hardening and motion, which are often prevalent in traditional CT and UTE MRI scans. Additionally, the synthetic CT images closely approximated the image quality of real CT

scans, making them a viable alternative for diagnostic purposes.
In the evaluation of structural alterations, such as bronchiectasis, wall thickening, and mucus plugs, synthetic CT images consistently outperformed UTE MRI, showing stronger correlations with real CT images. The synthetic CT also demonstrated a notable reduction in both false positives and false negatives, which is crucial for ensuring accurate clinical assessments. The Bhalla score analysis further supported the superior performance of synthetic CT, showing near-perfect concordance with real CT scores, particularly in the evaluation of CF-related pulmonary alterations.
However, one limitation of our current approach, which relies on 2D slice-wise image synthesis, is the difficulty in reconstructing anatomical structures accurately across different planes. As seen in Figure 2.11, the sagittal slices exhibit a staircase effect, a common issue with 2D reconstruction methods that fails to capture continuous spatial information along the third axis. To address this, future work could explore the use of 3D-based methods, which may improve the accuracy of spatial alignment and anatomical reconstruction by leveraging the full volumetric context of the data. In conclusion, while our study highlights the potential of synthetic CT generated from UTE MRI as a valuable radiation-free alternative for lung imaging, particularly in the management of CF and other pulmonary diseases, there remain areas for further exploration. The incorporation of 3D information could enhance image synthesis by addressing the limitations of 2D slice-wise methods, offering more comprehensive spatial information and further improving the precision of anatomical reconstructions.

Chapter Conclusion

This chapter presented a comprehensive study on the feasibility and optimization of MR to CT lung image synthesis using GANs. The initial objective was to explore and compare various generative models, both supervised and unsupervised, to assess their capability in generating high-quality synthetic CT images from UTE MRI data. The study was divided into several key sections, each addressing different aspects of the synthesis process, from the initial feasibility study to the optimization of preprocessing strategies, culminating in a clinical assessment of the generated synthetic images.
We initially explored the use of 2D GANs for MR to CT lung image synthesis, driven by the challenges associated with 3D methods, such as high memory costs and the need for an extensive dataset. By focusing on both paired and unpaired models, we found that the supervised model pix2pixHD, particularly when enhanced with Spatially-Adaptive Normalization (SPADE), consistently outperformed others. The SPADE method, with its adaptive normalization, helped to limit the hallucination of GAN features, resulting in synthetic CT images that closely resembled real CT scans. This approach underscores the significant potential of paired GAN models in achieving high-quality cross-modality image synthesis.
The critical role of data preprocessing was emphasized throughout the study, particularly the importance of resampling based on the modality with the highest resolution. By challenging state-of-the-art practices, we opted to use the higher resolution of the CT images to ensure the preservation of fine anatomical details, which are crucial for accurate image synthesis, especially in the lung regions. Additionally, the introduction of conditional registration using lung masks further improved the alignment between MR and CT images, contributing to better synthesis results.
The clinical assessment provided strong evidence that synthetic CT images generated from UTE MRI could potentially meet the standards required for clinical use. The synthetic images closely matched real CT images in terms of accuracy, particularly in the evaluation of cystic fibrosis (CF)-related structural alterations. This assessment also revealed that synthetic CT images produced fewer false positives and negatives compared to UTE MRI, reinforcing the viability of this approach as a radiation-free alternative for lung imaging.
However, the study also identified some limitations. While the synthetic images showed promise, the potential for GAN-generated images to introduce false positives or hallucinated features remains a concern, particularly in clinical contexts. Additionally, since the structures of interest, such as bronchi and vessels, are inherently 3D, the current 2D approach may not fully capture the spatial relationships and context necessary for accurate synthesis. This highlights the need to introduce 3D information to better leverage neighborhood data and improve the representation of complex anatomical structures. The reliance on standard image processing metrics, which may not fully capture the anatomical accuracy required in medical imaging, also underscores the need for more clinically relevant evaluation methods. Future research should focus on developing automated metrics that can reliably assess the accuracy of anatomical reconstructions, reducing the reliance on manual evaluations by expert radiologists.

Moving forward, in the next chapter we will continue to optimize 2D GANs, with a particular focus on refining loss function parameters, especially those related to perceptual loss, which can have a significant impact on the quality of reconstructions. Following this, we will start to introduce 3D information using 2.5D approaches to enhance spatial

coherence and more accurately capture complex lung structures with lightweight models. Additionally, a robust evaluation pipeline will be developed, integrating segmentation models to assess airway reconstruction, which will enable automated and consistent evaluation processes, thereby reducing dependency on expert radiologists and streamlining the validation of synthetic images for clinical use.

Chapter 3

Enhancing Synthesis with Perceptual Loss and 2.5D Approaches

The methods discussed in this chapter are related with the following publications :

[4] **Arthur Longuefosse**, Baudouin Denis de Senneville, Gaël Dournes, Ilyes Benlala, Pascal Desbarats, Fabien Baldacci. "On the use of perceptual loss for fine structure generation: Illustration on lung MR to CT synthesis". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2024)*.

[5] **Arthur Longuefosse**, Baudouin Denis de Senneville, Gaël Dournes, Ilyes Benlala, Pascal Desbarats, Fabien Baldacci. "Incorporating 3D information in 2.5D networks: strategies for MR to CT synthesis". In *IEEE 20th International Symposium on Biomedical Imaging, ISBI (2024)*.

In the previous chapter, we explored the feasibility of MR to CT lung image synthesis, focusing on various preprocessing optimizations, such as resampling, registration, and intensity normalization, to enhance the quality of generated images. We also demonstrated the efficacy of 2D GANs, particularly the paired model pix2pixHD with SPADE normalisation, in synthesizing high-quality CT images from UTE MRI data. While these initial steps provided a solid foundation, further refinement is necessary to improve the fidelity of the synthesized images, particularly in terms of preserving critical anatomical structures like vessels and bronchi.
In this chapter, we will first define task-specific metrics that allow us to quantitatively measure the reconstruction of fine structures such as vessels and bronchi, by leveraging segmentation networks in real and synthesized CT images. Following this, we will further optimize the synthesis process by refining the perceptual loss, a powerful approach that leverages high-level features from pre-trained networks to guide the generation of more realistic and anatomically accurate images. We will explore different strategies related to the use of perceptual loss in MR to CT synthesis, examining its impact on the quality of the synthesized images and identifying the most effective configurations.
Finally, we will introduce 3D information and context through the use of 2.5D approaches. This strategy allows us to retain the advantages of 2D implementations, such as the use of perceptual loss, which has been shown to significantly improve the synthesis process, while also incorporating crucial 3D spatial information. The 2.5D approach strikes a balance by maintaining a lightweight model with fewer parameters, enabling faster training

and inference, while still benefiting from the enhanced spatial coherence that 3D context provides.

3.1 Rethinking Evaluation Metrics For Medical Image Synthesis

In Section 2.3 of the previous chapter, we highlighted the critical need to accurately reconstruct fine anatomical structures, such as vessels and bronchi, in MR to CT lung image synthesis. Up until now, we have relied heavily on expert radiologists to quantitatively evaluate the quality of these synthesized images, carefully assessing the preservation of intricate anatomical details. However, this approach is both time-consuming and impractical for widespread application, as it requires significant manual effort and cannot be consistently employed in every study.
Recognizing these challenges, this section will explore the limitations of conventional metrics, which are commonly used in image processing and GAN-based synthesis. These traditional metrics provide a broad, global assessment of image quality but fail to capture the specific nuances needed to ensure that critical anatomical features are accurately preserved. To address this gap, we propose a task-specific approach that leverages segmentation networks to evaluate the reconstruction of anatomical structures in synthetic images. This method focuses on metrics that are robust to common issues such as registration errors and small spatial shifts, offering a more reliable and automated way to assess the quality of synthesized images in clinical contexts.

3.1.1 Limitations of Conventional Metrics

In the evaluation of cross-modality image translation, traditional image processing and GAN-based metrics have been widely used to assess the quality of synthesized images [42]. However, these metrics, while useful in certain contexts, fall short in capturing the specific needs of medical imaging, particularly when it comes to accurately reconstructing intricate anatomical structures such as vessels and bronchi [62, 53]

Intensity-based Metrics Traditional intensity-based metrics, such as Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM), are commonly employed to evaluate the fidelity of synthesized images. These metrics provide a measure of global image quality by comparing pixel intensities between the synthesized and reference images. For example, SSIM attempts to quantify image quality by considering luminance, contrast, and structural information, while MAE provides a simple average of absolute differences in pixel values.
These metrics can be useful in certain medical contexts, where accurate signal intensity is critical for dose planning. For example, in radiotherapy, MAE can be invaluable because precise reconstruction of signal intensities ensures that treatment plans are based on accurate tissue characterization. This precision is crucial for effectively targeting cancerous tissues while sparing healthy ones. Similarly, metrics like PSNR and SSIM are often employed in quality assurance processes to ensure that imaging systems produce outputs within acceptable tolerance levels, identifying deviations in image quality that might impact downstream analyses or treatments.
The SynthRAD2023 challenge [41], which focused on evaluating methods for pelvis and

brain MR to CT translation, exemplifies the reliance on such metrics in the field. In this challenge, synthesized images were evaluated using image processing metrics like MAE and SSIM. The best-performing methods in the challenge were those that optimized these metrics, resulting in images with intensities that closely matched the ground truth, a key factor in applications such as radiotherapy planning.
However, despite their utility in these specific applications, intensity-based metrics have significant limitations when applied to tasks such as lung MR to CT image synthesis:

- **Global evaluation** : They provide a broad assessment of image quality but do not account for the preservation of fine anatomical details. This is critical in medical imaging, where the accurate delineation of structures such as blood vessels, airways, and lesions is essential for diagnosis and treatment planning.

- **Sensitivity to noise and artifacts** : Metrics like SSIM and PSNR are sensitive to noise and low-level artifacts, which may not significantly affect the clinical utility of the images but can lead to misleadingly poor metric scores. This sensitivity can result in an inaccurate representation of the true quality and usability of the synthesized images.

- **Lack of clinical relevance** : These metrics do not consider the anatomical context, which is crucial for tasks such as surgical planning or disease monitoring. The inability to evaluate how well anatomical structures are preserved limits their effectiveness in medical image synthesis.

Moreover, optimizing metrics like MAE can lead to unintended consequences in image quality. As noted in previous studies on the use of L1 loss in synthesis models, optimizing for MAE often results in blurred images. This occurs because L1 loss constrains the network's optimization goal to the statistical average of all possible solutions within the task's solution space, leading to a smoothing effect that diminishes the sharpness of fine anatomical details [39, 98, 99].

GAN-based Metrics GAN-based metrics, such as the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), have been developed to evaluate the quality of generated images by comparing them to real images in a feature space. These metrics are particularly popular in the broader field of image synthesis due to their ability to measure the similarity between the distributions of real and synthesized images. Despite their widespread use in natural image synthesis, GAN-based metrics also exhibit significant limitations in the context of medical image synthesis:

- **Insensitivity to structural details** : Metrics like FID and KID primarly focus on global features extracted from pre-trained networks, which may overlook subtle but clinically important structural differences. Recent studies have shown that these metrics may not effectively capture global structural issues [100, 101].

- **Extrinsic properties** : GAN-based metrics rely on pre-trained networks that were typically trained on non-medical datasets. This can introduce biases and limit their applicability to the specific domain of medical imaging, where preserving anatomical accuracy is paramount.

Given these limitations, it becomes clear that traditional metrics, both intensity-based and GAN-based, are insufficient for fully evaluating the quality of synthesized medical images. They fail to capture the specific needs of medical imaging, particularly the accurate reconstruction of anatomical structures. To address this, we propose using task-specific metrics that are tailored to the particular objectives of the imaging task. This approach allows for a more meaningful evaluation, ensuring that the integrity of critical anatomical features is preserved in line with the initials goals of the synthesis process.

3.1.2 Introducing Task-Specific Metrics for Anatomical Reconstruction

In the pre-clinical study discussed in the previous chapter, we noted that the primary structures of interest in thoracic imaging, particularly for conditions such as mucoviscidosis, are the bronchi. These structures are crucial for defining the detection and severity of the disease, making their accurate reconstruction in synthesized images paramount. Building on this understanding, we developed a task-specific evaluation approach centered around bronchial segmentation. This method allows us to rigorously assess how well these critical structures are reconstructed in synthetic CT images compared to real CT scans, ensuring that our synthesis process meets the demands of clinical accuracy and relevance.

Segmentation Model

To achieve this, we employed the NaviAirway pre-trained network [102], a Bronchiole-sensitive Deep Learning-based Airway Segmentation Pipeline. This method has demonstrated its effectiveness in thoracic imaging, achieving the second highest Tree length Detected rate in the ATM22 challenge [103] during its long-term validation phase. Moreover, NaviAirway's implementation and pre-trained UNet weights are publicly available, making it a reliable and accessible choice for our evaluation purposes : https://github.com/AntonotnaWang/NaviAirway.
The NaviAirway model allows for the segmentation of bronchial structures from CT images, which we can be applied to our synthetic CT data to extract the bronchial tree. Since the actual ground truth bronchial tree is not available, we use the NaviAirway model to segment the real CT images, providing what we refer to as a *silver-standard* ground truth. This approach enables us to perform a comparative analysis between the segmentation results from synthetic CT and those from real CT, providing a robust framework for evaluating the anatomical accuracy of the synthesized images. An example of 3D airway segmentation from the NaviAirway model is presented in Figure 3.1.

Segmentation Metrics

To quantitatively assess the accuracy of the bronchial reconstructions in the synthetic CT images, we calculate segmentation metrics between the masks generated from the synthetic CT and the silver-standard ground truth derived from the real CT. The primary metric we use is the Dice score, an overlap metric that measures the similarity between the two segmentation masks. To ensure that the evaluation is focused solely on the bronchial structures within the lung, we constrain the Dice score calculation to the interior of the lung using the lungmask [90], effectively excluding the trachea from the analysis.
In addition to the Dice score, we employ the Normalized Surface Distance (NSD) [104]

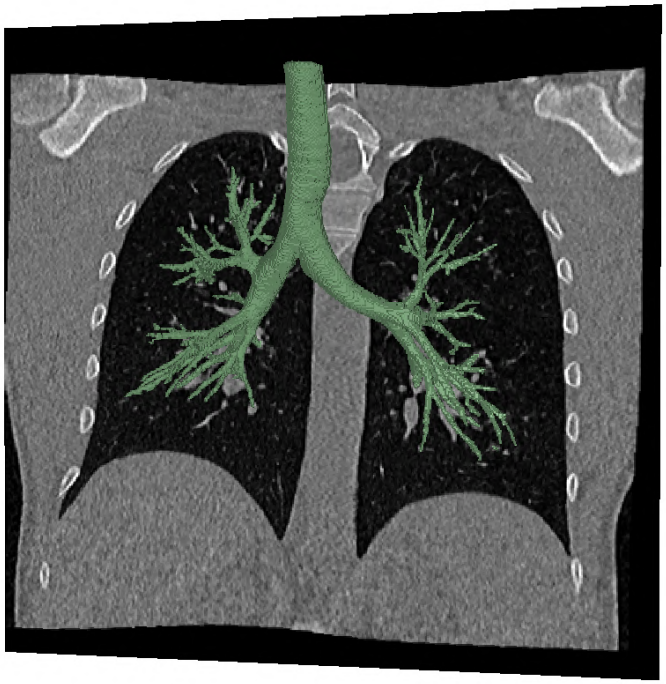


Figure 3.1: Example of CT airways segmentation using the NaviAirway pipeline.

metric, which offers a more nuanced evaluation of reconstruction quality, particularly for fine structures. The NSD metric is designed to be more robust to registration errors, providing a balanced assessment by considering surface deviations in the synthesized images [105]. Fundamentally, NSD can be viewed as an extension of the Dice score, with an added consideration for a manually defined margin of error. In our study, we set this margin as two times the voxel size, specifically 1.2mm, and, similar to the Dice score, constrain the NSD calculations within the CT lung mask. Figure 3.2 illustrates the computation of NSD, as explained in [105].

Together, these metrics provide a comprehensive evaluation of how well the bronchial structures are reconstructed in the synthetic CT images, ensuring that the synthesized images meet the anatomical accuracy required for clinical applications. By focusing on task-specific metrics, we can more effectively assess the performance of our synthesis models, ensuring that they align with the intended clinical objectives.
In addition to these task-specific metrics, our evaluation pipeline will also include traditional image processing metrics, such as SSIM, to provide a global assessment of image reconstruction quality. By combining these global metrics with our localized, task-specific metrics, we can achieve a balanced evaluation that captures both the overall fidelity of the images and the precise accuracy of critical anatomical structures. This dual approach allows us to ensure that our synthesized images not only look realistic on a broad scale but also faithfully reproduce the intricate details necessary for accurate clinical analysis.

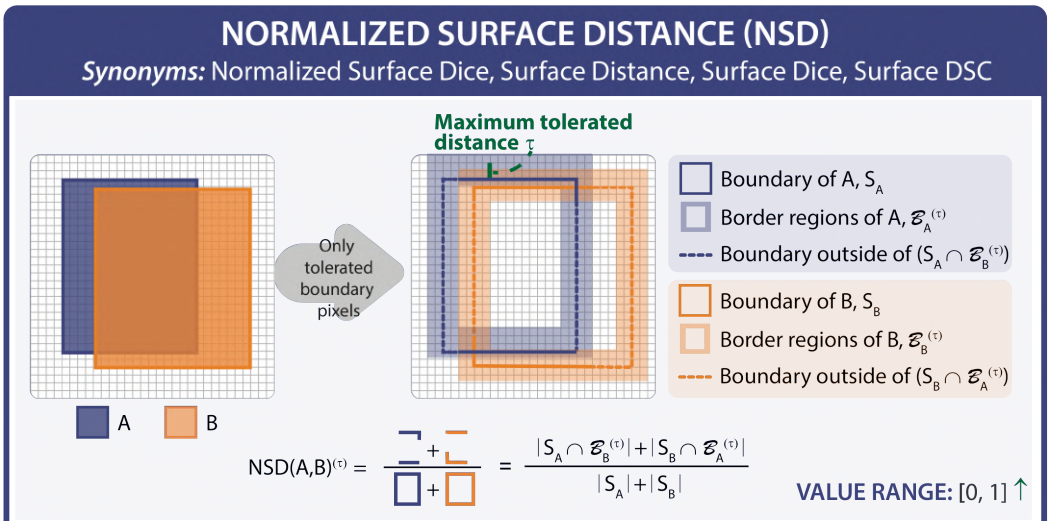


Figure 3.2: Illustration of the Normalized Surface Distance (NSD) metric. Reproduced with permission from the authors of the Metrics Reloaded website https://metrics-reloaded.dkfz.de.

3.2 Optimizing Perceptual Loss for Enhanced Synthesis

The evaluation metrics discussed in the previous section underscore the importance of accurately reconstructing fine anatomical structures, such as the bronchi, in synthesized CT images. However, ensuring that these structures are faithfully reproduced requires more than just robust evaluation methods, it necessitates the careful optimization of the synthesis process itself. One key aspect of this process is the loss function used during training, which directly influences how well the model can capture and preserve fine details in the generated images.

Traditional loss functions, such as pixel-wise comparisons (e.g., L1 loss), often fall short in this regard, leading to overly smooth and blurred images that lack the necessary detail for clinical relevance [39]. To address these challenges, perceptual loss has emerged as a powerful alternative, first introduced in 2016 [106]. This approach leverages high-level features extracted from pre-trained convolutional neural networks (CNNs) to guide the image synthesis process. By focusing on the perceptual similarity between the synthesized and real images, rather than just pixel-level accuracy, perceptual loss promotes the preservation of intricate textures and anatomical structures.

Previous research extensively explored perceptual loss for various image synthesis tasks, such as super-resolution [69, 106] and image-to-image translation [107]. However, many of these studies heavily relied on pre-trained VGG networks [108], initially trained on large-scale natural image datasets like ImageNet. This raises concerns regarding the transferability of learned features to medical imaging domains, where image characteristics and structures significantly differ from natural images.

In this section, we will explore the optimization of perceptual loss for thoracic MR to CT synthesis, with a particular focus on bronchial reconstruction. We will investigate various approaches, including the selection of network layers for feature extraction, the configuration of training datasets, and the impact of different data representations. By systematically analyzing these parameters, we aim to enhance the fidelity of the synthesized images, ensuring that the generated CT scans not only pass rigorous evaluation

metrics but also maintain the anatomical accuracy necessary for clinical applications.

3.2.1 Implementation Details

For our experiments, we used the pix2pixHD architecture with SPADE layers [44] as a baseline, building on the implementations detailed in Chapter 2, Section 2.2. This architecture facilitates image-to-image translation with spatially adaptive normalization, effectively addressing the generation of false positives that were observed in the standard pix2pixHD network [109]. The total loss function employed during training is a combination of adversarial loss ($\mathcal{L}_{adversarial}$ [25]), feature-matching loss ($\mathcal{L}_{FM}$ [43]), and perceptual loss ($\mathcal{L}_{perceptual}$ [106]).

$$\mathcal{L}_{total} = \mathcal{L}_{adversarial} + \mathcal{L}_{FM} + \mathcal{L}_{perceptual} \tag{3.1}$$

All experiments were conducted on the initial dataset of 110 patients, using the same optimized preprocessing techniques as described in Chapter 2, Section 2.2. This consistency ensures that the results are directly comparable to previous work and that any observed improvements can be attributed to the optimization of perceptual loss and other experimental parameters.

3.2.2 Exploration of Perceptual Loss Parameters

Perceptual loss is typically defined by calculating the mean absolute error (MAE) between the reconstructed image and the ground truth features extracted from one or multiple layers of a pre-trained network [106].

$$\mathcal{L}_{perceptual(x,y)} = \frac{1}{N}\sum_{i=1}^{N} \|\phi_i(x) - \phi_i(y)\|_1 \tag{3.2}$$

N being the number of layers extracted from the VGG network, ϕ_i the features of the i-th layer, x and y the synthesized and ground truth images to compare.

Data Configuration

The primary objective is to ensure compatibility between the input image size and the network architecture. Convolutional networks are designed to handle various image input sizes that are multiples of 2^n [110], where n represents the number of downsampling layers of the model. To evaluate the impact of input image sizes, we compare two approaches : resizing the input images to 224x224, as the original training set on ImageNet, and keeping the original input size of 512x512. While resizing the images brings them closer to the original training data distribution, it also introduces a loss of information due to interpolation.
In addition to image size, we investigated the effect of intensity normalization. Standard practice recommends applying the same normalization used during the pre-training of the network [111]. This typically involves z-score normalization of the images using the mean and standard deviation from the ImageNet dataset.

$$x_{norm} = \frac{x - mean_{ImageNet}}{std_{ImageNet}} \tag{3.3}$$

The results of our quantitative evaluation, presented in Table 3.1, highlight the importance of preserving the original image resolution. Models trained and evaluated with the full 512x512 input size consistently outperformed those using the resized 224x224 images, indicating that maintaining high-resolution is critical for accurate bronchial reconstruction. Furthermore, normalizing the input data to match the ImageNet mean and standard deviation values proved beneficial, with the normalized configuration (denoted norm.) achieving the highest Dice score (0.461 ± 0.035) and SSIM (0.800 ± 0.036) between the synthesized and real CT images. These findings underscore the significance of both image resolution and proper normalization in optimizing the effectiveness of perceptual loss for this task.

Table 3.1: Impact of data preprocessing on synthesis models using perceptual loss with VGG19 network. Reconstruction performance is assessed using image processing metrics (MAE, SSIM) and task-specific metrics (Dice, NSD) between real and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Loss	**MAE**	**SSIM**	**Dice**	**NSD**
VGG	65.95 ± 10.99	0.794 ± 0.024	0.438 ± 0.050	0.551 ± 0.075
VGG resized	69.16 ± 9.84	0.771 ± 0.017	0.423 ± 0.047	0.542 ± 0.064
VGG norm.	$\mathbf{65.80 \pm 10.17}$	$\mathbf{0.800 \pm 0.020}$	$\mathbf{0.467 \pm 0.051}$	$\mathbf{0.589 \pm 0.069}$

Layer Selection

When using a perceptual loss, it is common to consider only the activations before the fully connected layer at the end of the network. However, incorporating the activations from different layers throughout the network can provide valuable information about the image at various scales [43, 112]. Typically, low-level features, such as those captured by `relu0_1` and `relu1_3` at the beginning of the network, tend to capture edge and boundary information. On the other hand, high-level abstract features capture more complex shape and texture information [113]. By considering the activations from different layers, we can assess the impact of feature representation at various levels of abstraction. This allows us to evaluate the importance of different features for fine structure generation and helps determine the most effective layer selections for the synthesis process. Figure 3.3 shows an example of feature extraction in different layers in the ResNet50 network.
In addition to layer selection, we also compare different weightings assigned to the features extracted from each layer. For instance, we examine equal weights for all layers or adopt an ascending weight approach (e.g. $[\frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1]$) following state-of-the-art approaches [43] that aim to balance the weight of each layer depending on the number and spatial dimensions of features.
The results of our evaluation, presented in Table 3.2, reveal that the traditional approach of assigning the same weights to all layers consistently yielded the best performance. In contrast, the descending weights technique, commonly used in previous studies, led to inferior results. These findings underscore the importance of considering all layers equally in generating fine structures and ensuring overall visual quality. Moreover, our experiments show that excluding any single layer consistently results in a degradation of performance, further emphasizing the critical role that each layer plays in the pre-trained network when used for perceptual loss.

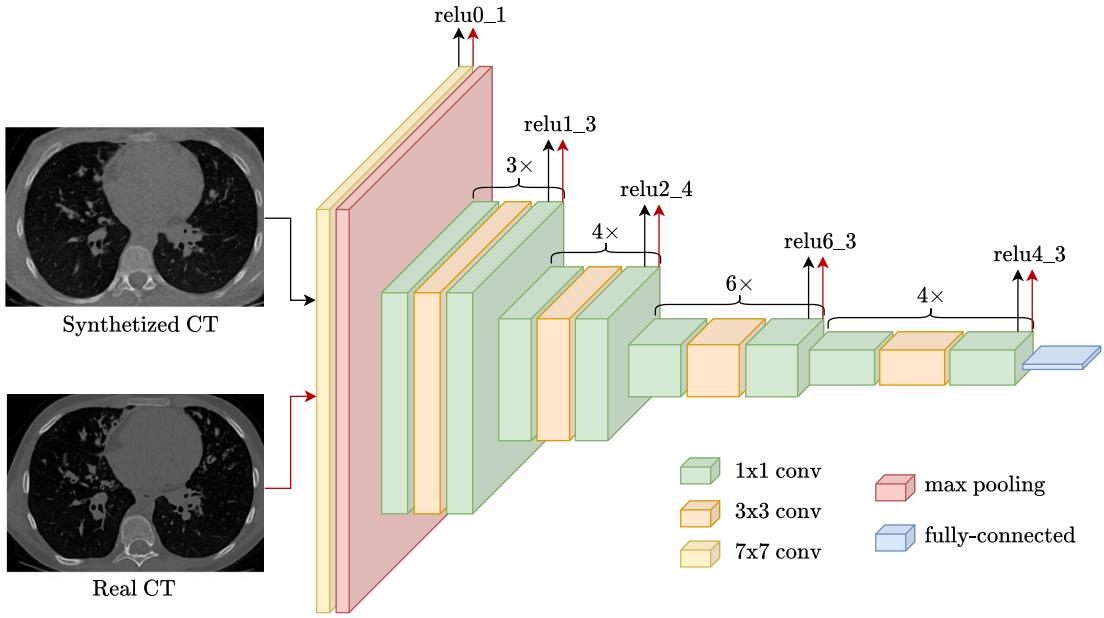


Figure 3.3: Example of feature extraction in different layers in the ResNet50 network. All features are acquired after the ReLU activation in each convolutional block.

Table 3.2: Comparison of different weights and layer selection in the VGG19 network, using image processing metrics (MAE, SSIM) and task-specific metrics (Dice, NSD) between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Layer strategy	**MAE**	**SSIM**	**Dice**	**NSD**
Descending weights	65.80 ± 10.17	0.800 ± 0.020	0.467 ± 0.051	0.589 ± 0.069
Same weights	**64.37 ± 9.74**	**0.829 ± 0.021**	**0.494 ± 0.059**	**0.614 ± 0.074**
Only first layer	65.49 ± 10.87	0.791 ± 0.020	0.439 ± 0.052	0.574 ± 0.065
W/o first layer	64.79 ± 11.68	0.809 ± 0.019	0.474 ± 0.055	0.592 ± 0.068
W/o middle layer	67.37 ± 11.46	0.760 ± 0.018	0.473 ± 0.056	0.590 ± 0.074
W/o last layer	68.46 ± 11.94	0.738 ± 0.023	0.492 ± 0.058	0.601 ± 0.076

For these evaluations, we maintained the ImageNet normalization method, as it demonstrated superior results in previous tests (Table 3.3). The results highlight the critical role of each layer in the VGG19 network, with the traditional approach of equal weighting across layers emerging as the most effective strategy for both fine structure generation and overall image quality.

Network Architecture

VGG [108] has been widely used as a backbone network for perceptual loss computation. It is known for its simplicity and effectiveness in capturing image features at different scales, and has demonstrated impressive performance in style transfer [114]. Additionally, we include ResNet50 [115] and ConvNeXt [116] in our comparison, as these two architectures have proven to be effective in various computer vision tasks. ResNet introduced residual connections, which improves convergence through a better gradient flow. ConvNeXt is a modernized ResNet based on grouped convolutions to effectively capture multi-scale information, surpassing previous CNNs in ImageNet classification tasks. By

evaluating these newer architectures alongside VGG, we aim to explore their potential advantages in terms of fine structure generation.

Table 3.3: Impact of data preprocessing on synthesis models using perceptual loss with VGG19 network. Reconstruction performance is assessed using image processing metrics (MAE, SSIM) and task-specific metrics (Dice, NSD) between real and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Loss	**MAE**	**SSIM**	**Dice**	**NSD**
L1 Loss	67.53 ± 10.31	0.813 ± 0.015	0.457 ± 0.056	0.573 ± 0.071
VGG	**64.37 ± 9.74**	**0.829 ± 0.021**	0.494 ± 0.059	0.614 ± 0.074
ConvNext Base	67.98 ± 10.74	0.824 ± 0.022	0.487 ± 0.066	0.607 ± 0.084
ResNet50	65.91 ± 10.24	0.820 ± 0.019	**0.504 ± 0.064**	**0.615 ± 0.082**

Our quantitative evaluation, presented in Table 3.3, shows that all three architectures, VGG19, ResNet50, and ConvNeXt, performed similarly in terms of reconstructing fine bronchial structures, with only minimal differences in the results. ResNet50 slightly outperformed VGG and ConvNeXt in fine structure reconstruction, achieving the highest airways Dice score and NSD. However, VGG led in overall image quality, as indicated by the highest MAE and SSIM score. ConvNeXt also delivered competitive performance, closely matching the results of VGG and ResNet50. These findings suggest that while newer architectures like ResNet50 and ConvNeXt offer incremental improvements, the choice of network architecture has a relatively minimal impact on the overall performance in the context perceptual loss for MR to CT synthesis.

Training Dataset

Pre-trained networks are commonly based on natural image datasets, such as ImageNet, which may not directly align with the medical imaging domain. However, recent advancements in the field have introduced large-scale medical image datasets, such as RadImageNet [117], which consists of over 1 million annotated images from various modalities. Pre-trained models based on this database have demonstrated promising results in transfer learning for medical imaging tasks [117]. In our study, we evaluate the impact of the training dataset on the generation of fine structures in medical images. We compare a ResNet50 trained on ImageNet, which can benefit from the diverse range of natural images and captured features, with a ResNet50 model trained on RadImageNet, which is more likely to have a comprehensive understanding of the characteristics present in medical images. In addition, we explore the influence of fine-tuning on the model performances.
In addition to comparing the training datasets, we explore the influence of fine-tuning the models. Fine-tuning involves unfreezing the weights of a pre-trained model and allowing them to be updated during training on a specific task, in this case, our MR to CT synthesis. This approach aims to adapt the pre-trained features more closely to the target domain, potentially improving performance.
The results of our evaluation, summarized in Table 3.4, indicate that the ResNet50 model trained on ImageNet (RN50-IN) outperformed the model trained on RadImageNet (RN50-RIN) in terms of both airways Dice score (0.504 ± 0.045) and SSIM (0.820 ± 0.030). This suggests that the large variety of images in ImageNet contributes to a more efficient

Table 3.4: Comparison of perceptual loss architectures using image processing metrics (MAE, SSIM) and task-specific metrics (Dice, NSD) between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Model	**MAE**	**SSIM**	**Dice**	**NSD**
RN50-IN	**65.91 ± 10.24**	**0.820 ± 0.019**	**0.504 ± 0.064**	**0.615 ± 0.082**
RN50-IN unfrozen	67.64 ± 9.42	0.804 ± 0.018	0.464 ± 0.058	0.566 ± 0.075
RN50-RIN	66.28 ± 9.80	0.785 ± 0.018	0.483 ± 0.061	0.589 ± 0.079
RN50-RIN unfroz.	69.65 ± 10.27	0.743 ± 0.017	0.456 ± 0.058	0.557 ± 0.074

capture of both low-level and high-level features, which are beneficial for MR to CT synthesis. Interestingly, models with frozen weights consistently performed better than their fine-tuned counterparts. This observation aligns with recent studies [118, 119], where it was found that fine-tuning can sometimes degrade performance, likely due to overfitting or the disruption of well-established feature representations that generalize well across different tasks.
These findings highlight the robustness of features learned from a large, diverse dataset like ImageNet, even when applied to medical imaging tasks, and suggest that fine-tuning may not always be beneficial, especially when it comes to maintaining the integrity of pre-trained features that generalize well to new tasks.

Contrastive Approach

Recent advancements in image-to-image translation have introduced the PatchNCE loss [52, 120], a bidirectional contrastive loss that extends the concept of traditional perceptual loss. While conventional perceptual loss relies on the similarity between features extracted from synthesized and target images, PatchNCE incorporates principles from contrastive learning to enhance the feature comparison process. Specifically, it leverages mutual information from embedded patches, contrasting the similarity between corresponding image patches with the dissimilarity of non-corresponding patches. This approach is similar to methods like SimCLR [121], which use contrastive loss to learn more robust and discriminative representations.
By incorporating these contrastive learning principles, PatchNCE offers several key advantages. First, it enhances feature learning: by contrasting positive and negative pairs, the model learns more discriminative and informative representations, leading to improved preservation of fine details in the synthesized images. Second, it improves training dynamics: the contrastive loss provides a smoother and more interpretable loss trajectory, facilitating easier evaluation of convergence and more efficient training.

Our evaluation, as shown in Table 3.4, demonstrates that both VGG and ResNet models using the contrastive approach outperformed their counterparts utilizing traditional perceptual loss. The VGG model with contrastive loss achieved an airways Dice score of 0.512 and an SSIM of 0.835, compared to 0.494 and 0.829 with traditional perceptual loss. Similarly, the ResNet model showed improvements, with the contrastive approach yielding better image quality and airway reconstruction scores. These results illustrate the effectiveness of the contrastive method in enhancing the preservation of fine anatomical structures.
Qualitative analysis of these results indicates that the contrastive approach improves both

Table 3.5: Comparison of perceptual loss architectures using image processing metrics (MAE, SSIM) and task-specific metrics (Dice, NSD) between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Model	**MAE**	**SSIM**	**Dice**	**NSD**
VGG perceptual	64.37 ± 9.74	0.829 ± 0.021	0.494 ± 0.059	0.614 ± 0.074
VGG contrastive	**63.12 ± 9.72**	**0.835 ± 0.018**	**0.512 ± 0.060**	**0.622 ± 0.081**
ResNet perceptual	65.91 ± 10.24	0.820 ± 0.019	0.504 ± 0.064	0.615 ± 0.082
ResNet contrastive	65.28 ± 10.13	0.822 ± 0.018	0.509 ± 0.063	0.617 ± 0.081

the VGG and ResNet models' ability to accurately reconstruct fine bronchial structures in the synthesized CT images. This demonstrates the effectiveness of incorporating contrastive learning principles into the perceptual loss framework. Notably, the best results were achieved using the contrastive loss on the VGG features, further reinforcing the minimal impact of model choice on overall performance, as observed in previous sections.

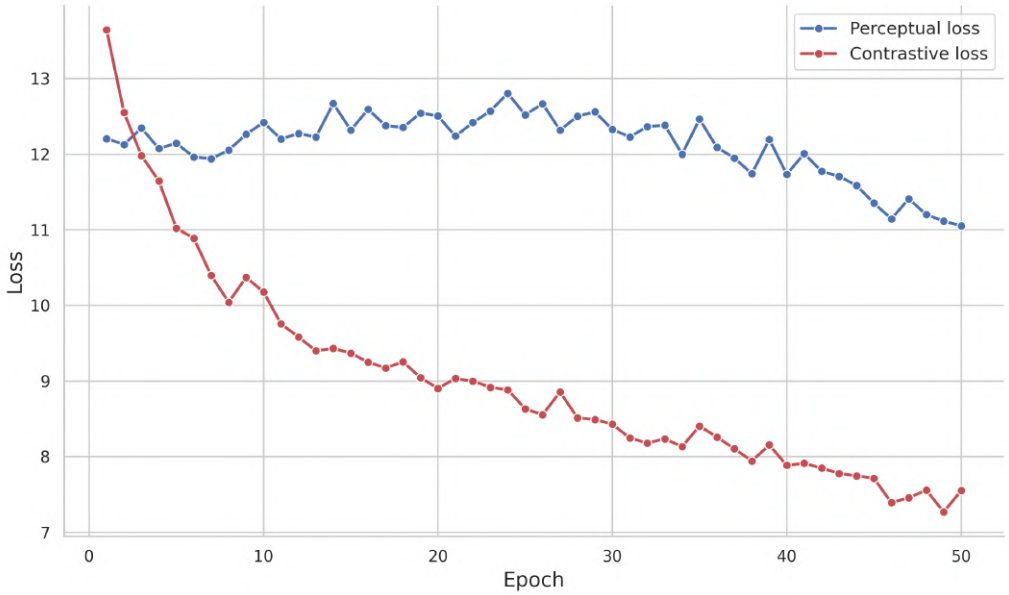


Figure 3.4: Comparison of VGG perceptual loss and VGG contrastive loss. The contrastive loss shows a faster and more pronounced convergence compared to traditional perceptual loss.

Figure 3.4 presents a comparison of the loss curves between the traditional VGG perceptual loss and the VGG contrastive loss. The contrastive approach not only leads to faster convergence but also results in a more pronounced separation of features, which aligns with the findings of recent studies [120]. This improved convergence behavior suggests that the contrastive loss facilitates more effective learning, allowing the model to better discriminate between relevant and irrelevant features, ultimately leading to higher quality image synthesis.

Qualitative Evaluation

Figure 3.5 shows a CT slice and the corresponding zoomed-in results generated by the models. The pixel-wise fake CT appears noticeably blurry, aligning with initial observa-

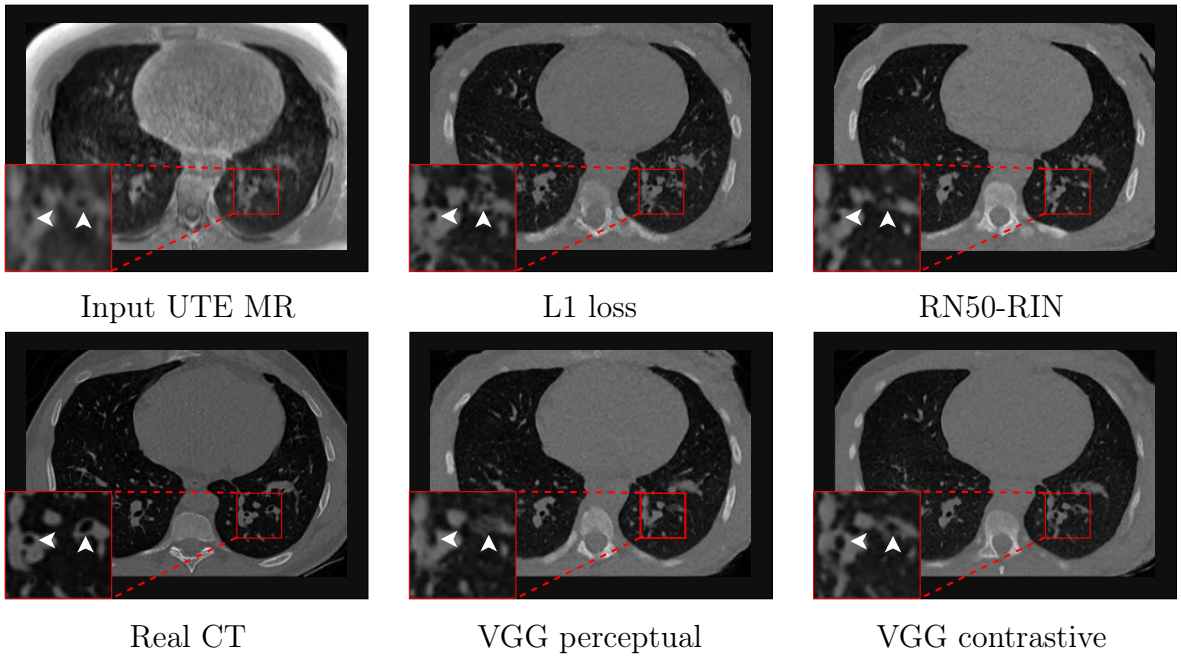


Figure 3.5: Comparison between input UTE MR, real CT and synthesized CT from model with various perceptual loss implementations. White arrowheads depict two bronchiectasis, which were closely reconstructed in the synthetic CT with contrastive loss only.

tions [39]. The ResNet50 model without the first layer presents poor boundary information and pronounced false positives in bronchi reconstruction, suggesting that the later layers of the network may prioritize large structure over local information. The VGG model with all layers demonstrates satisfying visual quality but falls short in the precise generation of structure, particularly for accurate bronchi delineation. The ResNet50 model trained on ImageNet yields the overall best results when employed as a perceptual loss. Leveraging the same weights for all layers of this model proves to be effective in capturing improved details for texture and structure generation, emphasizing the importance of network architecture and feature selection for the reconstruction of fine structure in MR to CT synthesis.

3.2.3 Synthesis of Findings

In this study, we explored various parameters influencing the effectiveness of perceptual loss in MR to CT synthesis, with a particular focus on bronchial reconstruction. By systematically analyzing the impact of preprocessing, network architecture, feature selection, and the introduction of a contrastive approach, we aimed to optimize the synthesis process and improve the fidelity of the generated images.

Impact of preprocessing First, we assessed the influence of image preprocessing before feeding them into the pre-trained network. Convolutional networks can process images of any size that is a multiple of 2^n, where n is the number of downsizing layers. This flexibility allowed us to compare two approaches: preserving the original image size and resizing images to match the size used during the training of the pre-trained network (e.g., 224×224 for ImageNet). The results in Table 3.1 demonstrate that maintaining the original dimensions of the images ensures that the network has access to all relevant infor-

mation, leading to better overall performance. On the other hand, normalizing the data to align with the statistical characteristics of the model further enhanced performance. This suggests that aligning the input data with the model's expectations can improve its capability to accurately generate fine structures.

Network architecture Among the architectures tested, the ResNet50 model emerged as the top performer compared to VGG and ConvNeXt models (Table 3.3), both in terms of image quality and fidelity in reconstructing bronchial structures. This superiority can be attributed to ResNet's improved gradient flow, which enables better information propagation and stability compared to VGG. Although the ConvNeXt model achieved satisfactory results overall, the reduced resolution in each layer, caused by the stride 4 convolutions, may have negatively affected the reconstruction of fine details. Additionally, the ResNet50 model pre-trained on ImageNet outperformed the one pre-trained on RadImageNet, suggesting that the extensive variety and diversity of images in ImageNet play a significant role in effectively capturing both low-level and high-level features. Moreover, fine-tuning the model proved to be less effective than leaving the weights frozen. In the context of perceptual loss, fine-tuning or employing a medical dataset for pre-training seems unnecessary since the pre-trained network serves as a fixed feature extractor, already providing valuable and generalizable properties. As outlined in previous studies [122], this may be because the convolutional filters learned from medical images exhibit less prominent edges, which could limit the effectiveness of fine-tuning in tasks requiring precise reconstruction of fine structures.

Feature selection Initially, we hypothesized that edge details would be sufficient for generating fine structures, leading us to believe that low-level features would outperform higher-level features. However, the results showed that using only the first layer led to a significant decrease in performance, indicating that low-level features alone are insufficient for this task. Interestingly, removing the last layer also resulted in a slight decrease in performance, suggesting that high-level features are crucial for fine structure generation as well. Therefore, we conclude that a combination of low-level and high-level features is essential for effectively synthesizing fine structures in MR to CT synthesis. Furthermore, we found that using uniform weighting across selected layers was sufficient, eliminating the need for introducing a hyperparameter for individual layer weights. This simplification reduces the complexity of the model and avoids the challenges associated with fine-tuning layer-specific weights.

Contrastive Loss Incorporating a contrastive approach to perceptual loss, specifically through the use of PatchNCE, brought significant improvements in both the training process and the quality of the synthesized images. Computing perceptual loss in a contrastive manner facilitated faster and more stable convergence of the training model. Quantitatively, this approach led to better results in both global image quality and local reconstruction of airway structures, as evidenced by the higher Dice scores, NSD and SSIM values compared to traditional perceptual loss. The contrastive method's ability to emphasize the similarity between corresponding patches while distinguishing non-corresponding ones proved effective in enhancing the preservation of fine anatomical details, further validating the utility of this advanced loss function in medical image synthesis.

This study evaluated the impact of various parameters on the effectiveness of perceptual loss in MR to CT synthesis, with a particular emphasis on bronchial reconstruction. We found that maintaining the original image resolution and aligning input data with the statistical properties of pre-trained models led to better overall image quality and structural accuracy. ResNet50, especially when pre-trained on ImageNet, proved to be the most effective architecture for this task. Additionally, introducing a contrastive approach to perceptual loss provided further improvements in feature representation and model convergence.
However, this study also highlights certain limitations. Notably, our implementation of perceptual loss is confined to 2D frameworks, leveraging large-scale models like VGG and ResNet that were originally trained on natural image datasets such as ImageNet. This 2D approach may not fully capture the three-dimensional complexity inherent in medical imaging, particularly for structures like the bronchi that extend across multiple planes.
To address this limitation, the next phase of our research will explore the incorporation of 3D information through a 2.5D approach. This method strikes a balance between the simplicity and computational efficiency of 2D models and the spatial coherence offered by 3D data. By integrating 3D context while retaining the advantages of 2D perceptual loss, we aim to further enhance the fidelity and clinical relevance of our synthesized CT images.

3.3 Bridging the Gap Between 2D and 3D Synthesis : Strategies for 2.5D

In the current landscape of medical imaging, emerging deep learning technologies dominate research across various data modalities [113]. The tasks within medical imaging, including segmentation, classification, and synthesis, often exhibit common features, with advancements in one task potentially benefiting others [123]. A significant debate in this field is the integration of 3D information in networks when handling volumetric medical images, posing a choice between 2D and 3D approaches.
Traditional 2D methods face limitations due to a lack of neighborhood information, which can lead to suboptimal reconstructions in alternate axes and disconnection between structures across slices [124]. Conversely, full-3D approaches not only face significant memory challenges, limiting their ability to process entire volumes, but they also tend to require more data to achieve robust performance. While patched-3D methods strive to balance data quantity and information content, they suffer from issues like a lack of long-range dependencies and edge effects on patch borders during the reconstruction process. Moreover, all 3D approaches, whether full-volume or patch-based, are challenged by an increased number of parameters, requiring more data, supervision, and posing greater training complexities [125, 126].
To overcome these challenges, 2.5D methodologies have emerged as a promising middle ground [127, 128]. These strategies aim to incorporate 3D contextual information while maintaining the advantages of 2D methods, such as lower computational demands and the ability to utilize pre-trained 2D networks.
In this section, we will begin by providing an overview of existing 2.5D methods, highlighting their implementations and benefits in medical imaging tasks. Following this, we will conduct a comparative study between 3D and 2D convolutions in 2.5D approaches, exploring how the convolution computation impact the quality of synthesized images. Building

on these insights, we will propose a hybrid approach that combines the strengths of 2.5D techniques to enhance MR to CT thoracic image translation. Finally, we will present the results of our experiments, offering insights into how these strategies can be optimized to achieve more accurate and coherent reconstructions in the context of thoracic imaging. Through this exploration, we aim to demonstrate the potential of 2.5D approaches in bridging the gap between 2D and 3D synthesis, ultimately improving the clinical relevance of our synthetic images.

3.3.1 Overview of Existing 2.5D Methods

To effectively incorporate 3D contextual information while maintaining the computational efficiency of 2D networks, various 2.5D strategies have been proposed in medical imaging. These methods primarily utilize 2D convolutional networks, which are often criticized for representing 3D information in a channel-wise manner [124, 129]. This approach can result in the initial convolution merging depth information, potentially leading to the loss of valuable 3D structural details, as illustrated in Figure 3.7.
In our work, we focus on combining two prominent 2.5D techniques, multi-axial fusion and thickened slice, to introduce neighborhood information without the computational drawbacks associated with full 3D reconstruction, and capitalizing on the advantages of perceptual loss [106]. We then compare these methods with 2D and 3D convolutions to discern the effectiveness of using 2D convolutions to convey 3D information. Finally, we evaluate our approach within the context of MR to CT synthesis in lungs, employing a robust evaluation based on structural similarity, task-specific metrics based on fine structure reconstruction, and a comparison of the number of parameters, memory cost, and training time for each method. This comprehensive evaluation provides valuable insights on the use of 2.5D networks in terms of performances and computational resources on volumetric medical images. Figure 3.6 illustrates the state-of-the-art 2.5D architectures.

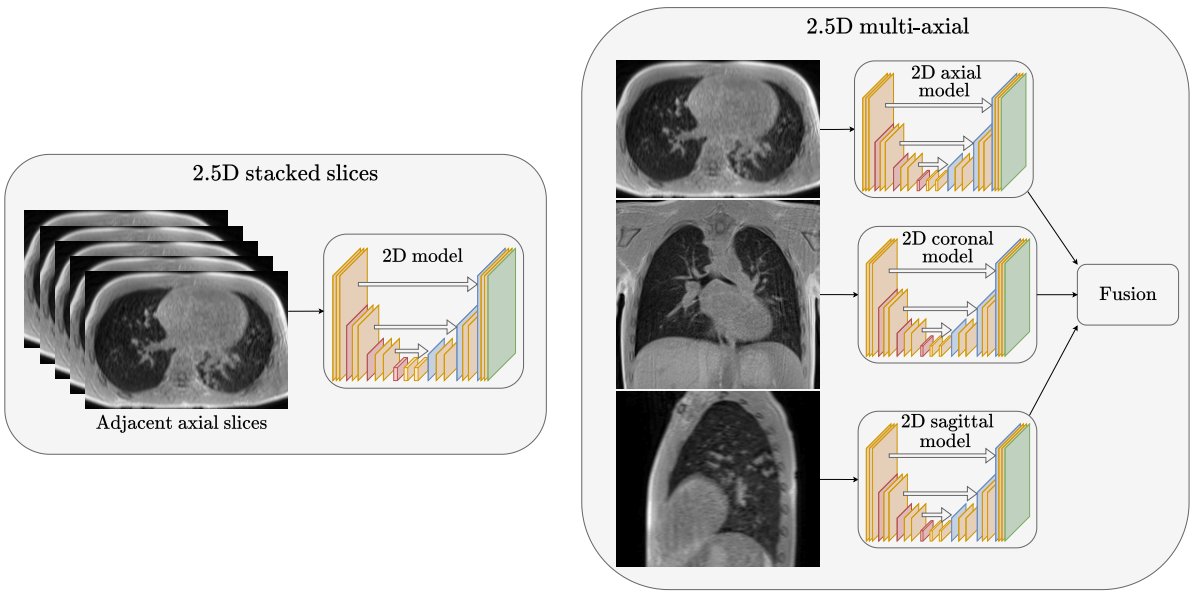


Figure 3.6: Comparison of 2.5D architectures. Stacked slice approach, where multiple adjacent 2D slices are combined as input, and multi-axial fusion, where separate networks process slices from axial, coronal, and sagittal planes, followed by fusion to capture 3D spatial context

- 2.5D Multi-axial: This approach involves training separate 2D networks on slices from different anatomical planes, such as the axial, coronal, and sagittal views. After training, the results from each network are fused, typically by voting, median aggregation, or employing an additional CNN. For instance, Prasoon et al. applied this method to the segmentation of knee MRIs, improving the accuracy of cartilage boundary detection across multiple planes [130].

- 2.5D Thickened Slices: In this method, a series of neighboring 2D slices are stacked together to create a multi-channel input that captures local 3D context. For example, a stack of 5 neighboring slices (5x512x512) is processed as a single input by a 2D network [124]. Applications of thickened slices have been successfully demonstrated in medical imaging, such as in MS lesion segmentation, where adjacent slice information enhanced the detection accuracy [131].

Despite the strengths of these various 2.5D strategies, their reliance on 2D convolutions, particularly in methods like thickened slices, can lead to premature merging of depth information, which may prevent the accurate preservation of fine 3D anatomical structures. In the following sections, we will compare several implementations of 2.5D approaches for MR to CT synthesis, focusing on their strengths and limitations. By highlighting the trade-offs of these methods, we propose a novel hybrid approach that combines the benefits of multiple 2.5D strategies to enhance synthesis performance and ensure better structural coherence in volumetric medical imaging.

3.3.2 Implementation Details

For this section, we build upon the best settings identified in previous work to establish a robust baseline for our experiments. We continue to use the pix2pixHD architecture [43] enhanced with SPADE normalization layers [44], as detailed in Chapter 2. This combination has been shown to effectively handle image-to-image translation tasks, particularly in addressing issues like false positives, as observed in earlier experiments.
In addition to this architectural foundation, we incorporate the optimized perceptual loss settings identified in section 3.2. Specifically, we employ the VGG model pre-trained on the ImageNet dataset, integrated with a contrastive learning approach. This setup has demonstrated superior performance in preserving fine anatomical structures, such as bronchi, in MR to CT synthesis tasks.
The baseline for our experiments consists of a 2D implementation of this model. For each of the tested 2.5D approaches discussed in the upcoming sections, specific adaptations are made to this baseline model to accommodate the unique requirements of each approach. These modifications are detailed within each method's respective section.
All experiments were conducted using the same dataset of 110 patients and adhered to the optimized preprocessing pipeline established in Chapter 2, ensuring consistency and comparability across all experiments.

3.3.3 2.5D Multi-Axial Strategies: Fusion of 2D Networks

A prevalent strategy in 2.5D networks involves employing multiple 2D networks across different anatomical planes, such as axial, coronal, and sagittal viewsproach leverages the strengths of 2D convolutional networks while incorporating 3D context by processing the data from multiple orientations. Each 2D network operates independently along one axis,

and the results are subsequently fused to produce a final synthesized volume.
For segmentation tasks, methods like majority voting are often used [132], where each plane's prediction is considered, and the most frequent label is selected for the final output. In contrast, synthesis tasks may utilize more advanced fusion strategies, such as adding an additional CNN to process the fused results, aiming to refine and improve the quality of the synthesized volume. However, the latter introduces challenges, including increased parameter tuning and the need for an additional loss configuration, which can complicate the training process. Simpler fusion methods, such as taking the mean or median of the results from each axis, are often preferred due to their ease of implementation and robustness.
For our implementation, this 2.5D multi-axial approach does not require significant modifications to the baseline models presented in Section 3.3. Instead, we trained the baseline 2D model three times, each time focusing on a different anatomical plane: axial, coronal, and sagittal. After training, we reconstructed the 3D synthesized volumes for each model by stacking the 2D slices along their respective axes. The final multi-axial volume was obtained by calculating the median of the corresponding voxels from the three volumes generated from each axis. This median fusion method was chosen for its robustness to outliers, which is particularly beneficial in maintaining the structural integrity of fine anatomical details.

Table 3.6: Comparison of 2D and 2.5D models using image processing metrics (MAE, SSIM) and task-based metrics (Dice, NSD) using the airways segmentation from NaviAirway, between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Model	**MAE**	**SSIM**	**Dice**	**NSD**
2D axial	63.12 ± 9.72	0.835 ± 0.018	0.512 ± 0.060	0.622 ± 0.081
2D coronal	64.03 ± 10.26	0.833 ± 0.017	0.507 ± 0.057	0.613 ± 0.078
2D sagittal	63.88 ± 9.85	0.827 ± 0.020	0.511 ± 0.058	0.624 ± 0.075
2.5D multi-axial	**60.84 ± 10.54**	**0.840 ± 0.016**	**0.537 ± 0.058**	**0.661 ± 0.074**

As shown in Table 3.6, the performance of the individual 2D models along the axial, coronal, and sagittal planes was consistent, with all three achieving similar scores across the evaluated metrics. The real advantage of the multi-axial approach becomes apparent when these individual models are combined. By applying a median fusion of the synthesized volumes from the three planes, the 2.5D multi-axial strategy achieved a noticeable improvement in both image quality and the accuracy of anatomical reconstructions. The SSIM score slightly increased, indicating a better overall similarity between the synthesized and real CT images. More importantly, the task-specific metrics around airways reconstruction saw a substantial improvement, with the Dice score increasing to 0.504 in the multi-axial approach, and the NSD also showing a corresponding enhancement.
This increase in performance suggests that the multi-axial fusion method effectively leverages the complementary information captured in different planes, leading to more accurate and reliable reconstructions. The simplicity of this approach is particularly noteworthy, as it relies solely on training multiple 2D models independently, without requiring complex modifications or additional computational resources. The demonstrated efficacy of this method underscores the value of 2.5D multi-axial strategies in enhancing the quality of medical image synthesis, offering a practical and efficient alternative to more complex methodologies.

3.3.4 To 3D or not 3D : That is the Convolution

In 2.5D approaches, one common method to incorporate 3D contextual information into a 2D framework is by stacking multiple neighboring 2D slices along the channel dimension, creating an input volume with dimensions such as 5x512x512. This technique, often referred to as *thickened* or *stacked* slices, introduces a degree of spatial coherence that single 2D slices lack. However, a significant limitation arises when traditional 2D convolutional layers are employed within the network architecture. The initial convolutional layers effectively merge the 3D information into a 2D format by treating the stacked slices as separate channels, potentially leading to the loss of valuable depth information (see Figure 3.7). This can undermine the network's ability to accurately capture and reconstruct fine anatomical structures that extend across multiple slices.

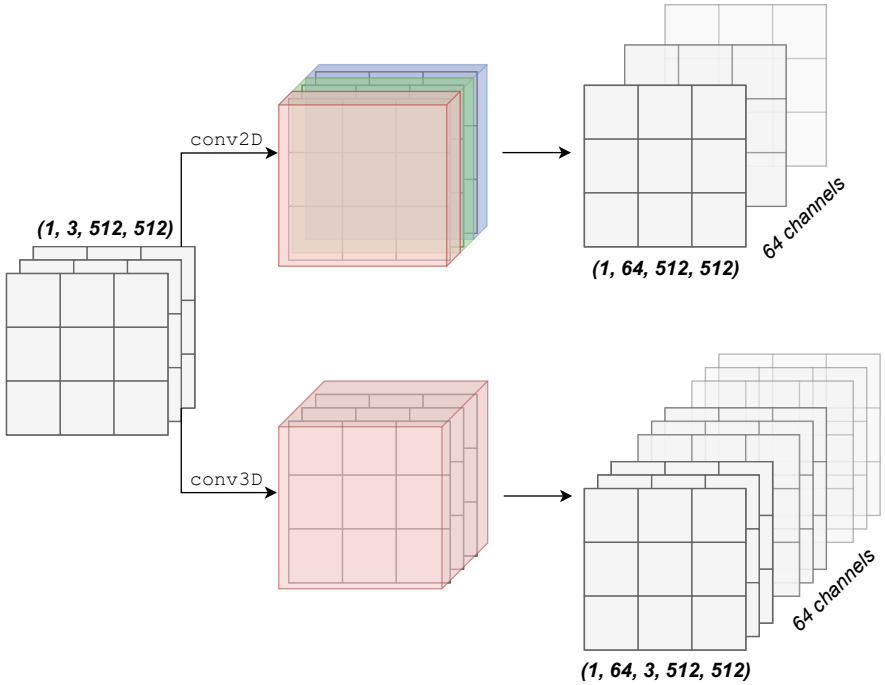


Figure 3.7: Comparison of 2D and 3D convolution kernels on a 2.5D stacked slices of size 3x512x512. In the `conv2D` approach, the depth information appears to be fused in the channel dimension after the first convolution [124, 131], while the `conv3D` approach preserves this information by keeping a dedicated depth channel.

To address this limitation, we propose an alternative approach that replaces the traditional 2D convolutions (`conv2D`) with 3D convolutions (`conv3D`) when processing stacked slices. By employing 3D convolutions, the network can maintain and utilize the depth information throughout the learning process, potentially improving the accuracy of the synthesized images, particularly for tasks requiring the reconstruction of complex 3D structures, such as MR to CT translation.
In our study, we implemented the pix2pixHD architecture with SPADE normalization layers, leveraging both 2D and 3D configurations to compare their performance in the 2.5D stacked slices context. To adapt the SPADE layers for use with 3D convolutions, we modified the downsampling components of the generator and the SPADE normalization CNN. Specifically, we applied a stride of 1 for the depth dimension while keeping a stride of 2 for the width and height dimensions. This adjustment ensures that the slice

thickness is preserved throughout the convolutions, allowing the network to maintain consistent depth information while downsampling the other dimensions.
The results of our evaluation between the traditional `conv2D`-based approach and our proposed `conv3D`-based method are summarized in Table 3.7. The comparison between the two approaches reveals that while the overall image quality, as measured by SSIM and MAE, remains similar between the `conv2D` and `conv3D` configurations, the use of 3D convolutions offers a slight improvement in the reconstruction of fine anatomical details. The Dice score, which assesses the overlap between the segmented airways in the synthesized and real CT images, shows a modest increase from 0.549 to 0.564 when using 3D convolutions. Similarly, the NSD, which evaluates surface reconstruction accuracy with an allowance for slight registration errors, improves from 0.693 to 0.723.

Table 3.7: Comparison of 2.5D stacked slices approaches using image processing metrics (MAE, SSIM) and task-based metrics (Dice, NSD) using the airways segmentation from NaviAirway, between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Model	**MAE**	**SSIM**	**Dice**	**NSD**
2.5D stacked `conv2D` axial	60.12 ± 10.81	0.843 ± 0.018	0.544 ± 0.059	0.663 ± 0.075
2.5D stacked `conv3D` axial	**59.67 ± 11.24**	**0.845 ± 0.020**	**0.551 ± 0.061**	**0.671 ± 0.082**

These results suggest that the 3D convolutions in the stacked-slices approach enhance the network's ability to preserve and utilize 3D information, leading to more accurate reconstructions of fine structures. This is particularly significant in medical image synthesis tasks, where the accurate depiction of 3D anatomical features is crucial for clinical applications.

3.3.5 Proposed Hybrid 2.5D Approach

To further enhance the performance of MR to CT synthesis, we propose a novel hybrid 2.5D approach that integrates the strengths of the multi-axial and stacked slices strategies. This approach aims to create a more comprehensive and powerful network capable of leveraging both local and global 3D information. By deploying multiple 2.5D networks along each anatomical axis (axial, coronal, and sagittal), and utilizing the stacked slice method to maintain depth information within each network, the hybrid model combines multi-directional data with a robust preservation of 3D structures. This fusion allows the network to capture rich, multi-scale features that are crucial for accurate and coherent image synthesis. Figure 3.8 presents an illustration of our proposed hybrid approach.
For the implementation of the hybrid 2.5D approach, we conducted experiments using both the `conv2D` and `conv3D` methods within the stacked slices framework. This comparison allows us to assess whether the 3D convolution approach, which preserves depth information more effectively, has an impact when combined with the multi-axial strategy. Specifically, we trained multiple 2.5D networks, each corresponding to a different anatomical axis, and then fused the outputs using median aggregation to produce the final synthesized volume.
This hybrid model was designed to balance the benefits of each individual approach, leveraging the spatial coherence from multi-axial views and the depth preservation from

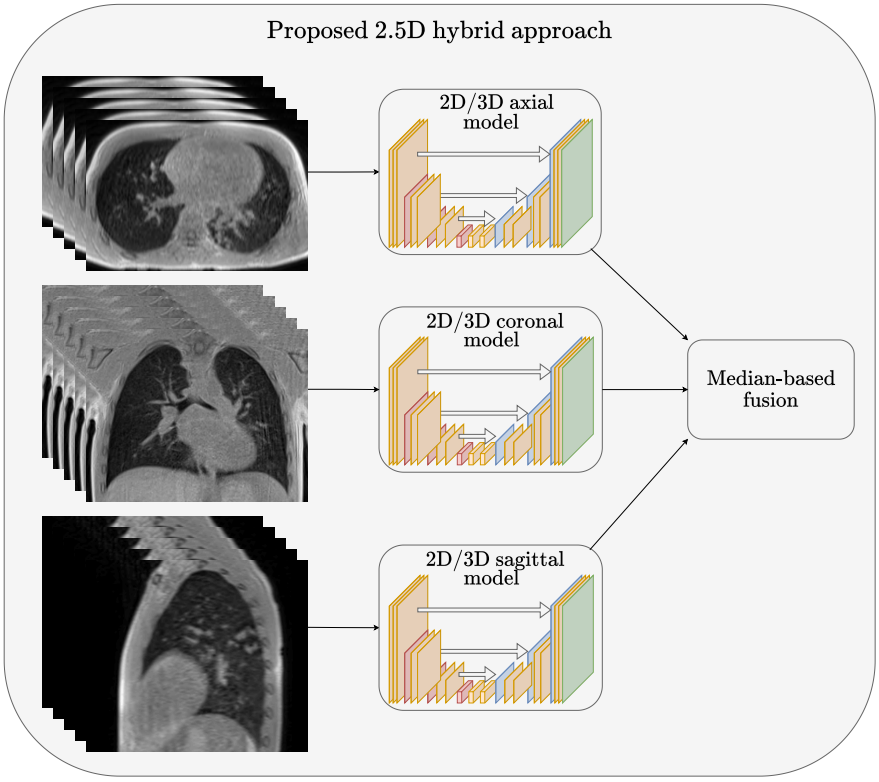


Figure 3.8: Illustration of the proposed hybrid 2.5D approach. Each axis is processed by a separate network using either 2D or 3D convolutions within the stacked slice framework.

stacked slices, thus aiming to achieve superior results in both image quality and the reconstruction of fine anatomical details.

We compared the visual quality of the reconstructed 3D volumes produced by each network, as a qualitative evaluation. Figure 3.9 illustrates the coronal views across all models, highlighting their differences in retaining 3D information during synthesis.

The 2D network, trained on axial slices, exhibited a noticeable staircase effect in the coronal view, indicating its limitations in capturing multi-dimensional information. The multi-axial 2.5D network reduced this effect by incorporating information from all three anatomical planes, but the resulting output appeared slightly blurred due to the fusion process. In contrast, the 2.5D stacked slices model showed significant improvements in sharpness, successfully recovering finer anatomical details, although minor residual artifacts remained. Overall, the proposed hybrid approach, which combines the strengths of multi-axial and stacked slice methods, delivered the best results. It achieved superior smoothness, accuracy, and sharpness across all views, making it the most effective solution for retaining both spatial coherence and depth preservation during synthesis.

The performance of the hybrid 2.5D approach, along with the individual results from previous sections, is summarized in Table 3.8, with a quantitative evaluation using intensity-based metrics and task-specific metrics based on airways reconstruction.

The results indicate that the hybrid approach delivers the overall best performance across all metrics. In terms of image quality, as measured by MAE and SSIM, the hybrid approach slightly outperforms other methods, achieving the highest SSIM score of 0.852. This suggests that the combined strategy effectively enhances the perceptual quality of the synthesized images.

However, the most significant improvements are observed in airway reconstruction. The

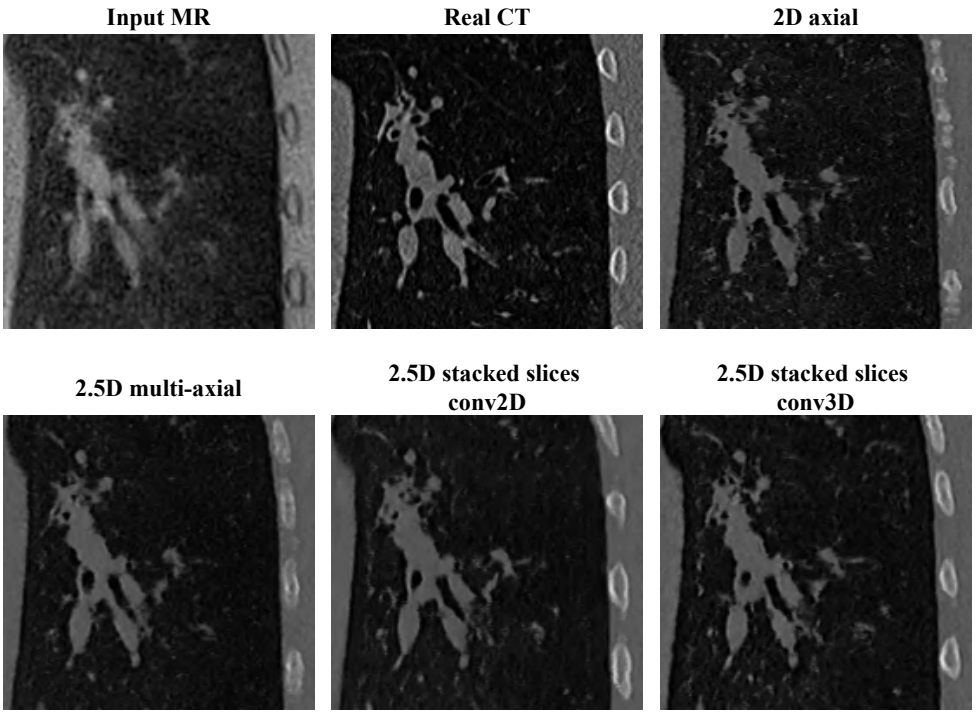


Figure 3.9: Comparison of 2D and 2.5D MR to CT synthesis models. Slices are zoomed in on the left lung and presented in coronal view.

Table 3.8: Comparison of 2D and 2.5D models using image processing metrics (MAE, SSIM) and task-based metrics (Dice, NSD) using the airways segmentation from Navi-Airway, between real CT and synthesized CT images.

	Intensity-based metrics		Task-based Metrics	
Model	**MAE**	**SSIM**	**Dice**	**NSD**
2D axial	63.12 ± 9.72	0.835 ± 0.018	0.512 ± 0.060	0.622 ± 0.081
2.5D multi-axial	60.84 ± 10.54	0.840 ± 0.016	0.537 ± 0.058	0.661 ± 0.074
2.5D stacked `conv2D` axial	60.12 ± 10.81	0.843 ± 0.018	0.544 ± 0.059	0.663 ± 0.075
2.5D stacked `conv3D` axial	59.67 ± 11.24	0.845 ± 0.020	0.551 ± 0.061	0.671 ± 0.082
2.5D hybrid `conv2D`	57.39 ± 10.45	0.851 ± 0.017	**0.572 ± 0.062**	0.695 ± 0.080
2.5D hybrid `conv3D`	**57.13 ± 10.24**	**0.852 ± 0.017**	0.568 ± 0.063	**0.698 ± 0.076**

Dice score for the 2.5D hybrid `conv2D` approach reached 0.572, indicating a substantial improvement in the overlap between synthesized and real CT airways. The NSD metric, which assesses the accuracy of surface reconstructions with an allowance for registration errors, also showed significant gains, with the best performance achieved by the 2.5D hybrid `conv3D` approach at 0.698.
Interestingly, the discrepancy between the `conv2D` and `conv3D` stacked slices approaches, which was more pronounced in the single-axis tests, is reduced when these methods are combined with the multi-axial strategy. This suggests that the potential loss of information associated with using 2D convolutions can be mitigated by incorporating multi-axial

data, which provides additional spatial context and improves the overall reconstruction quality.

3.3.6 Computational Efficiency

The computational costs of each method are summarized in Table 3.9 and reveal a clear trade-off between performance and resource consumption. The 2.5D stacked slices method with `conv2D` adds minimal computational overhead compared to the 2D baseline. This modest increase in memory cost and parameters, combined with an acceptable rise in training time per iteration, makes it a computationally efficient option. The primary reason for this additional cost is the need to compute perceptual loss over the additional slices used in the stacked approach. Despite the longer per-iteration time, the overall training time remains comparable due to the larger volume of images processed in each iteration.

Table 3.9: Comparison of 2D and 2.5D networks computational resources in terms of memory cost on GPU, number of parameters and training time per iteration.

Model	**Memory Cost**	**Parameters**	**Time per iter.**
2D axial	6.5 GB	92.0 M	0.150 s
2.5D stacked `conv2D`	7 GB	92.1 M	0.280 s
2.5D stacked `conv3D`	16 GB	274.5 M	1.8 s

On the other hand, the `conv3D` approach significantly increases both memory cost and training time per iteration. This reflects the additional complexity of 3D convolutions and the exponential growth in the number of parameters. While `conv3D` achieved the highest performance, the substantial increase in computational costs raises questions about its feasibility, particularly in resource-constrained environments or when training on larger datasets. Given the relatively minor performance improvements of `conv3D` over `conv2D`, it may not justify the additional costs in all scenarios.

3.3.7 Synthesis of Findings

In this study, we explored and compared several 2.5D approaches for MR to CT synthesis, aiming to integrate 3D information into networks traditionally designed for 2D processing. By employing techniques such as multi-axial fusion and stacked slices, we sought to bridge the gap between 2D and 3D methods, with a particular focus on reconstructing fine anatomical structures like airways. Figure 3.10 illustrates the segmentation results for airways across different approaches, showing the 2D-based synthetic CT, 2.5D stacked slice synthetic CT, and the real CT for reference.
Each approach was assessed in terms of both its effectiveness in enhancing airway reconstruction and its computational efficiency. The results demonstrate the potential of 2.5D approaches in balancing reconstruction accuracy with manageable resource demands, while highlighting the strengths and limitations of each method.
The multi-axial fusion approach, which trains 2D networks across different anatomical planes (axial, coronal, and sagittal), effectively incorporated multi-directional information, improving both image quality and airway reconstruction. Fusion of outputs from these planes led to significant improvements in reconstruction metrics such as Dice and NSD. However, the slight loss of sharpness observed in the qualitative results indicates

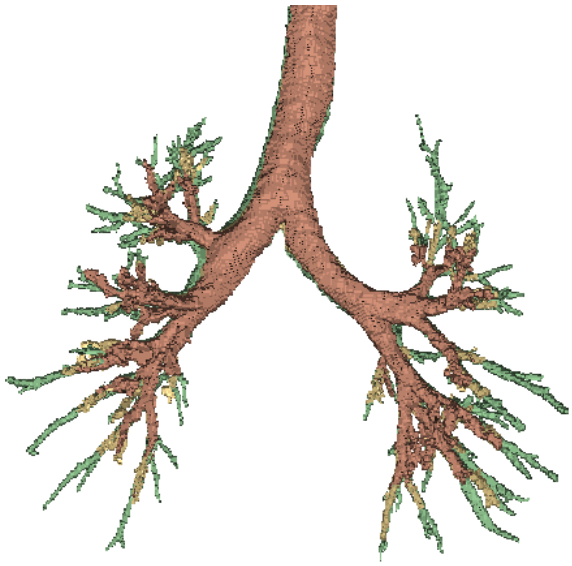


Figure 3.10: Comparison of airways segmentations from 2D synthetic CT (red), 2.5D stacked slices `conv2D` synthetic CT (yellow) and real CT (green) using the NaviAirway pipeline [102].

that blending outputs across planes, while capturing 3D context, may come at the cost of fine detail preservation.
The stacked slices method introduced depth information by stacking adjacent slices before processing. This approach enhanced airway reconstruction over the baseline 2D method, preserving 3D structure without significant loss of detail. The use of `conv2D` provided a good balance between performance and efficiency, with only minor increases in memory usage and training time. While `conv3D` delivered marginally better results, the gains were not substantial enough to justify its higher memory and processing costs, indicating that `conv2D` can still capture relevant 3D features effectively.
The hybrid approach, which we propose as a novel combination of the multi-axial and stacked slice methods, emerged as the best-performing strategy for airway reconstruction. By training multiple stacked-slice networks along each axis, this method successfully leveraged the strengths of both approaches, introducing multi-directional information while maintaining the 3D context provided by the stacked slices. This hybrid method achieved the highest Dice and NSD scores, reflecting its effectiveness in reconstructing fine anatomical structures. Qualitatively, it delivered the most accurate and detailed reconstructions while mitigating both the staircase effect and blurriness seen in other approaches.
Despite the promising results, certain limitations remain. Full 3D approaches, while potentially offering the most complete spatial context, may not be feasible with our current dataset due to the significant increase in memory requirements and computational complexity. However, patch-based 3D approaches could be explored in future studies to strike a balance between 3D information retention and resource efficiency. One key challenge is the lack of 3D perceptual losses, as most are based on large, pre-trained 2D models like VGG, which are trained on natural image datasets such as ImageNet. This limits the ability to apply perceptual loss in a fully 3D context. Future work will focus on developing 3D perceptual loss models or implementing similar topological constraints that can better capture the complexity of 3D medical data, enabling a more effective use of 3D convolutions and improving overall synthesis accuracy without overwhelming computational demands.

Chapter Conclusion

In this chapter, we explored and optimized various approaches for MR to CT synthesis, with a particular focus on enhancing bronchial reconstruction through advanced evaluation metrics, perceptual loss, and 2.5D strategies. First, we re-evaluated traditional image processing metrics, highlighting their limitations in medical image synthesis, especially for the reconstruction of fine anatomical structures like airways. To address this, we introduced task-specific metrics such as the Dice score and Normalized Surface Distance (NSD), which were tailored to assess the quality of airway reconstructions using the NaviAirway segmentation pipeline. These task-based metrics provided a more clinically relevant evaluation framework, complementing traditional image quality metrics.
Next, we optimized the use of perceptual loss for 2D MR to CT synthesis by systematically investigating the impact of different parameters, including preprocessing techniques, network architecture, and layer selection. By focusing on fine anatomical details, we demonstrated that ResNet50 pre-trained on ImageNet, combined with contrastive perceptual loss, significantly improved the quality of the synthesized images. This approach facilitated better preservation of bronchial structures, particularly when compared to traditional loss functions. However, we identified the limitation that current perceptual loss is confined to 2D frameworks, which may not fully capture the 3D complexity inherent in medical imaging.
Finally, we introduced and compared several 2.5D approaches to incorporate 3D information into networks traditionally designed for 2D processing. The multi-axial fusion strategy and stacked slices approach both demonstrated improved image quality and airway reconstruction compared to the 2D baseline, but the most effective method was our proposed hybrid approach. By combining multi-axial fusion with stacked slices, we achieved the best performance in reconstructing fine anatomical structures. The hybrid method allowed for the integration of 3D information without the significant computational burden of full 3D models. Although the use of `conv3D` slightly improved results over `conv2D`, the increase in memory cost and complexity was disproportionate to the gains, making `conv2D` the more practical choice for most applications.
Looking ahead, the next chapter will build upon these findings by exploring scalable 3D architectures for medical image translation. We will also develop new loss functions tailored for the reconstruction of anatomical structures, leveraging segmentation model embeddings to parallel the concept of perceptual loss in 3D. Additionally, the evaluation pipeline will be enhanced to include segmentation of multiple anatomical features, allowing for a more precise, task-specific assessment of the overall synthesis quality across various structures.

Chapter 4

Advanced 3D Synthesis with Scalable Architectures and Anatomical Guidance

---

The methods discussed in this chapter are related with the following publications :

[2] **Arthur Longuefosse**, Baudouin Denis de Senneville, Gaël Dournes, Ilyes Benlala, Fabien Baldacci, Pascal Desbarats. "Anatomical feature-prioritized loss for enhanced MR to CT translation". Under revision in *Computers in Biology and Medicine*.

[7] **Arthur Longuefosse**, Edern Le Bot, Baudouin Denis De Senneville, Rémi Giraud, Boris Mansencal, Pierrick Coupé, Pascal Desbarats, Fabien Baldacci. "Adapted nnU-Net: A Robust Baseline for Cross-Modality Synthesis and Medical Image Inpainting". In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI), 2024, Held in Conjunction with MICCAI 2024*.

In the previous chapter, we focused on enhancing MR to CT synthesis by developing task-specific metrics, such as the Dice score and Normalized Surface Distance (NSD), to locally assess the reconstruction of fine structures like airways, offering a more clinically relevant evaluation alongside traditional image quality metrics. We also investigated the optimization of perceptual loss, demonstrating the contrastive approach on embeddings trained from ImageNet significantly improved the fidelity of bronchial reconstructions by capturing fine anatomical details more effectively than traditional loss functions. Finally, we integrated 3D information through 2.5D strategies, such as multi-axial fusion and stacked slices, which allowed for a balance between 2D simplicity and 3D spatial coherence, with the hybrid approach yielding the best results for airway reconstruction.
In this final chapter, we will advance towards a truly scalable 3D synthesis framework that leverages robust, reproducible methods for medical image translation. Our first objective is to adapt the nnU-Net architecture, a well-established framework in medical image segmentation, to the task of cross-modality synthesis. The nnU-Net has consistently proven to be a robust and highly generalizable architecture across numerous segmentation challenges, and by adapting it to MR to CT synthesis, we aim to harness its scalability,

reproducibility, and ease of use. This adaptation will allow us to address the complexities of 3D data while maintaining efficient training and inference times, offering a scalable solution for medical image translation.
In the second section, we will refine the synthesis process by introducing a novel 3D loss function that incorporates anatomical guidance. By utilizing features extracted from pre-trained medical segmentation models, we will ensure that critical anatomical structures, such as intra-lung lesions and extra-lung regions, are accurately reconstructed during synthesis. This anatomical-aware loss function will prioritize the preservation of medically significant structures, moving beyond traditional pixel-level losses.
To validate our approach, we will extend the task-based evaluation pipeline developed in the previous chapter, employing new segmentation networks to evaluate the reconstruction of a broader range of anatomical structures. This comprehensive evaluation will ensure that our method not only achieves high-quality synthesis globally but also maintains the anatomical fidelity required for clinical applications.

4.1 Robust and Scalable 3D Synthesis with nnU-Net

In this section, we explore how the nnU-Net framework [93], originally developed for medical image segmentation, can be adapted for 3D synthesis tasks such as cross-modality translation and image inpainting. We begin by introducing the core features of the nnU-Net framework and its success in medical imaging. Next, we explore the specific adaptations made to the architecture to enable robust 3D synthesis, detailing changes in preprocessing, architecture, losses, and volume reconstruction strategies. Following that, we validate the effectiveness of the adapted nnU-Net using public datasets and challenges for tasks such as pelvis MR to CT translation and brain inpainting. Finally, we reflect on the limitations of this approach in the context of our original task of lung MR to CT synthesis, discussing potential future improvements.

4.1.1 Introduction to nnU-Net Framework

nnU-Net (no-new-U-Net) [93] is a well-established framework in the field of medical image segmentation. Developed to address the complexity and variability in deep learning model design, nnU-Net self-configures for various medical segmentation tasks, eliminating the need for manual hyperparameter tuning. It has emerged as a robust and consistent baseline in many segmentation challenges, frequently outperforming more complex models, including those based on modern transformer architectures, despite its simplicity as a CNN-based approach. Recent studies have confirmed nnU-Net's ability to deliver competitive results across a wide range of medical datasets, establishing it as a reliable tool for segmentation tasks in medical imaging. [133, 134]
The nnU-Net framework is particularly known for its ability to adapt to different imaging modalities and anatomical regions, adjusting key hyperparameters such as patch size, network depth, and batch size based on the characteristics of the dataset. This adaptability, coupled with its standardized configuration, has led to its widespread adoption and consistent performance in medical imaging competitions and real-world applications. Figure 4.1 presents an overview of the nnU-Net pipeline.
Over the past decade, medical image synthesis has emerged as a significant trend in healthcare. This technique addresses the shortfall in medical datasets by producing synthetic data, thereby enhancing data diversity and improving the robustness of machine learning models. It also plays a crucial role in translating between different imaging modalities, which is beneficial for applications such as attenuation correction and radiation therapy planning. Additionally, medical image synthesis can significantly enhance patient comfort by reducing the need for repeated imaging procedures and minimizing exposure to harmful radiations. The significant advancement in this field is mainly attributed to the evolution of generative methods, such as generative adversarial networks [25] (GANs) and diffusion models [32] (DMs) . Numerous recent studies have explored synthetic data generation and cross-modality translation, addressing reconstruction challenges in anatomical areas such as the brain[48], pelvis [63], lungs [135].
In the specific area of image reconstruction, inpainting methods are also gaining interest for restoring missing or damaged parts of images in a plausible manner. This is especially relevant in brain MRI analysis, where traditional methods fail due to a lack of contextual information about disease-related anomalies. The introduction of challenges such as the BraTS Local Synthesis of Healthy Tissues [136] in 2023 underscores the growing emphasis on using inpainting to synthesize medical data. In practice, most state-of-the-

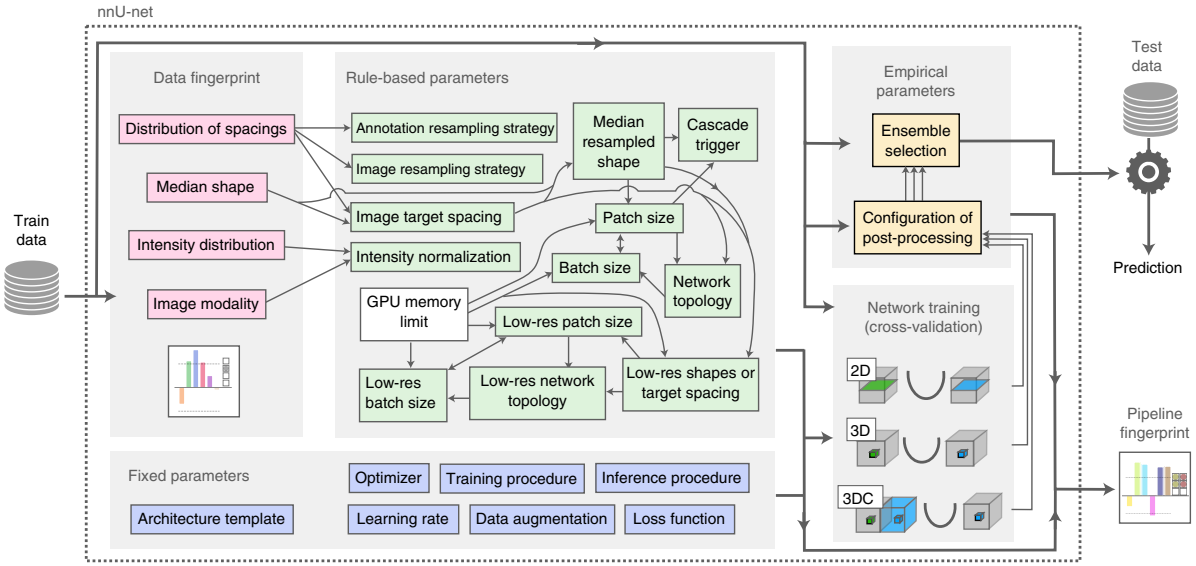


Figure 4.1: Overview of the automated nnU-Net method. Reproduced with permission from Springer Nature. Isensee et al., nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation [93].

art inpainting methods on natural images rely on diffusion models [137]. However, the heavy computational requirements of diffusion models and the processing of large 3D images necessitate custom strategies to manage memory requirements. For example, some propose using a 2D denoising model applied slice-wise on the axial plane [138]. Despite their effectiveness, diffusion models are hindered by the slow forward diffusion process, making them less suitable for time-constrained applications. Consequently, the primary approach for medical image inpainting remains Generative Adversarial Networks (GANs). For example, a 3D pix2pix [39] model was proposed as a baseline for the inpainting task in the BraTS2023 challenge.

Despite these advances, there remains a need for a robust and adaptable baseline in the fields of cross-modality translation and medical image synthesis. nnU-Net has already demonstrated its efficacy in the SynthRad2023 challenge [41] for medical image translation, ranking among the top methods. In this chapter, we present an adaptation of nnU-Net for 3D medical image synthesis, utilizing its self-configuring properties to create a flexible and powerful baseline for cross-modality translation and inpainting. Our adapted nnU-Net was tested on two specific tasks: Pelvis MR to CT translation and brain inpainting, using publicly available datasets from recent medical image challenges. The results demonstrate the potential of nnU-Net as a reliable baseline for 3D synthesis, offering a solution that balances performance, scalability, and ease of implementation.

To facilitate broader use and promote reproducibility, the adapted nnU-Net code has been made freely available on GitHub: https://github.com/Phyrise/nnUNet_translation.

4.1.2 Adapting nnU-Net for 3D Synthesis

In this section, we outline the key adaptations made to the nnU-Net framework to enable its application to 3D medical image synthesis tasks. These adjustments focus on preprocessing, architectural changes, loss function modifications, and strategies for reconstructing volumetric data.

Preprocessing

To adapt the preprocessing pipeline for image synthesis, we restructured the nnU-Net workflow, which is traditionally designed for (image, segmentation) pairs, to handle (image, image) inputs. Specifically, we disabled the cropping step in preprocessing because, by default, the cropping settings are not uniform across different modalities. For the remaining steps, we adhered to the nnU-Net guidelines [93]. This includes z-score normalization for MR images and CT normalization for CT images, where intensity values are collected from the foreground classes, the mean and standard deviation are computed, values are clipped to the 0.5 and 99.5 percentiles, and then normalized by subtracting the mean and dividing by the standard deviation. Finally, we adapted the data augmentation pipeline of nnU- Net. This pipeline includes a variety of spatial transformations, such as rotation, scaling, and elastic deformations, as well as intensity-based augmentations, such as brightness and contrast adjustments and the addition of Gaussian noise.

Architecture

We retained the core architecture of nnU-Net due to its robustness and flexibility. However, given the shift from segmentation to image synthesis, we made several key adaptations to optimize the architecture for generating continuous intensity values in the synthesized images in a floating point format, rather than discrete labels cast to integers.
Additionally, we added some variants for the decoder part of the network. In standard U-Net architectures, transposed convolution (also known as deconvolution) is typically used for upsampling in the decoder. However, recent studies have shown that when combined with perceptual loss, transposed convolutions can produce checkerboard artifacts in image synthesis tasks [139]. To address this, we implemented a combination of upsampling layer followed by convolution, which has been proven to be an effective alternative [140]. This method helps mitigate the appearance of artifacts and produces smoother outputs. We provided two types of interpolation for the upsampling: trilinear and nearest neighbor, which can be used as variants during training depending on the requirements of the task.

Loss Function

We adapted the loss function to better suit the requirements of image synthesis, shifting away from the segmentation-focused Dice loss. Specifically, we replaced it with mean absolute error (L1) loss and added the support of mean squared error (L2) loss.
A key addition to our framework is the integration of 3D perceptual loss. Recent work on large-scale 3D medical classification networks, particularly the MedicalNet model, has shown promise for extending the concept of perceptual loss into the 3D domain. MedicalNet was originally developed for transfer learning in medical imaging tasks, utilizing pre-trained features to improve performance across various medical datasets. By leveraging its ability to capture complex volumetric structures, MedicalNet provides a robust feature space that is more suitable for guiding 3D image synthesis.
Moreover, 3D perceptual loss using MedicalNet has been successfully applied in recent studies within the MONAI framework, particularly in image synthesis tasks, demonstrating its effectiveness in preserving anatomical details and enhancing overall synthesis quality [141].
We implemented support for 3D perceptual loss based on MedicalNet embeddings, which can be used either as a standalone reconstruction loss or in combination with L1 loss to

balance global intensity matching with perceptual accuracy.

Volume reconstruction

By default, nnU-Net processes images in patches with a 50% overlap in the inference phase. Traditionally, majority voting is used for segmentation tasks to merge overlapping patches. We replaced this approach with mean and median reconstruction methods for image synthesis.
Mean reconstruction, which averages the overlapping patches, tends to produce smoother and more globally consistent outputs. While this method often results in slightly blurrier images, it yields superior performance on global intensity-based metrics such as mean absolute error (MAE) and mean squared error (MSE). This makes it ideal for applications where pixel-wise accuracy is crucial, and small deviations across the entire volume must be minimized.
In contrast, median reconstruction was implemented to preserve high-frequency details, making the output visually sharper and more detailed. This method is particularly useful for tasks where preserving fine anatomical structures is important, such as in complex areas of the brain or lungs. However, because it focuses on retaining local details, median reconstruction tends to underperform on global intensity metrics compared to mean reconstruction. This trade-off between visual accuracy and global intensity metrics allows for flexibility in choosing the appropriate reconstruction strategy based on the specific needs of the task at hand.

4.1.3 Validation using Public Challenges and Datasets

In the previous section, we detailed the adaptations made to the nnU-Net framework to enable its use in 3D medical image synthesis, including modifications to the preprocessing pipeline, architecture, loss functions, and volume reconstruction methods. These adaptations aim to enhance the nnU-Net's capabilities for tasks like cross-modality translation and image inpainting, allowing it to generate high-quality synthesized medical images.
To validate this framework, we have chosen to evaluate it on two distinct tasks using public datasets from challenges in the pelvic and brain regions. These datasets provide a standard reference for MR to CT translation and local brain inpainting, which allows for direct comparison with the current state-of-the-art methods in these fields. By leveraging public challenges, we aim to facilitate the adoption of our framework by the broader community and ensure that our approach is robust and comparable to leading techniques. In this section, we present the datasets and tasks evaluated, followed by a detailed analysis of the results, including a comparison with methods from the respective challenges, as well as an evaluation of the computational costs.

Datasets

Pelvis dataset For the synthesis of pelvic images, we use the open-access dataset from Task 1 of the SynthRAD2023 challenge [41]. This dataset includes paired MR and CT images from 180 patients used for training, with an additional 30 patients designated for validation, with no ground truth CT publicly available. The dataset provides high-quality, well-aligned MR and CT image pairs, which eliminates the need for additional registration processes. The voxel sizes in this dataset are standardized to $2.5 \times 1 \times 1$ mm, ensuring

consistent spatial resolution across all images. The results of this MR to CT translation task were obtained using the SynthRad2023 challenge online validation server.

Brain dataset The 2023 BraTS Local Synthesis of Healthy Brain Tissues Training dataset is comprised of 1251 T1w images pre-processed with bias field correction, spatial resampling to $1 \times 1 \times 1$ mm, and cropped to a final spatial resolution of 155 x 240 x 240. For each image, a corresponding healthy and non-healthy tissue mask is provided. Non-healthy tissue mask, delineates areas affected by tumoral invasion, while healthy tissue mask are generated according to the process described in [136]. The dataset originally included a number of images for validation and testing, hosted on the synapse servers, however, evaluation on those images is no longer available since the closure of submission queues. As such, 251 subjects were randomly sampled and kept for testing. Remaining subjects were used for training.

Implementation details

For validation, we compared our adapted nnU-Net framework with a 3D version of pix2pixHD [43], a well-established GAN architecture. The 3D pix2pixHD model employed a patch-based approach similar to nnU-Net, with mean reconstruction on overlapping voxels.
Our adapted nnU-Net followed the original guidelines [93], with models trained for 1,000 epochs and 250 iterations per epoch. The patch sizes were automatically set to $64\times128\times224$ for the pelvis task and $96\times160\times160$ for the brain inpainting task. Training was performed using an NVIDIA RTX 6000 GPU with 24GB VRAM, allowing a batch size of 2.
The 3D pix2pixHD model was trained following the procedure in [43], with 200 epochs and 1,000 iterations per epoch. A fixed patch size of $128\times128\times128$ was used for both the pelvis and brain tasks, with training on the same hardware setup.
All models were implemented in PyTorch 2.3.0 with automatic mixed precision for computational efficiency. The training duration ranged from 24 to 48 hours, depending on the task. An 80/20 training/validation split was maintained for both datasets. For the pelvis MR to CT translation task, we report Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Peak Signal to Noise Ratio (PSNR). For the brain inpainting task, we use Mean Squared Error (MSE), SSIM, and PSNR. These metrics, while intensity-based, are standard in the respective challenges and allow for direct comparison with other state-of-the-art methods.

Pelvis MR to CT translation

The pelvis MR to CT translation task in the SynthRAD2023 challenge is conceptually similar to our initial lung MR to CT translation task. However, there are notable differences, particularly in the resolution and the type of information captured by each modality. In this dataset, both MRI and CT images share the same voxel size and spatial resolution. Furthermore, MRI provides more detailed information regarding soft tissues, such as organs and muscles, while CT scans are more suited for visualizing bone structures.
To validate our approach, we compared the performance of our adapted nnU-Net against a 3D version of the pix2pixHD model. In addition to comparing nnU-Net with standard MAE (L1 loss), we also explored the integration of perceptual loss using MedicalNet embeddings. This allows us to assess whether incorporating more high-level anatomical feature-based loss functions can further enhance synthesis accuracy.

Table 4.1: Comparison of 3D networks performances using MAE, SSIM, and PSNR metrics between real CT and synthesized CT for the pelvis MR to CT translation task. Evaluation is performed on 30 patients.

Model	MAE	SSIM	PSNR
3D nnU-Net MAE	**65.74 ± 12.72**	**0.851 ± 0.030**	**27.94 ± 1.48**
3D nnU-Net MedicalNet	90.48 ± 16.46	0.811 ± 0.032	26.69 ± 1.47
3D pix2pixHD	88.57 ± 17.53	0.815 ± 0.029	26.64 ± 1.33
Results from Synthrad2023 challenge report			
1st - SMU-MedVision			
3D Transformer U-Net	**58.83 ± 13.41**	**0.885 ± 0.029**	**29.61 ± 1.79**
2nd - Jetta_Pang			
3D nnU-Net MSE	65.73 ± 13.75	0.869 ± 0.032	28.38 ± 1.68
3rd - FAYIU			
3D Swin U-NetR	61.72 ± 13.32	0.876 ± 0.030	28.83 ± 1.61

Table 4.1 presents a quantitative evaluation of the adapted nnU-Net's performance based on MAE, SSIM, and PSNR metrics between synthesized and ground truth CT images. All metrics for our models are calculated using the SynthRAD2023 challenge validation server on 30 patients from the validation set. This table shows that using a standard 3D pix2pixHD model yields inferior performance compared to the nnU-Net for cross-modality synthesis. Results from the challenge report [42] are also listed for reference, although they are based on the test set and not on the validation set. The trends between nnU-Net with MAE and nnU-Net with MSE are strongly correlated. The challenge ranking is based not only on reported intensity metrics but also on dose evaluation metrics, explaining the observed ranking.

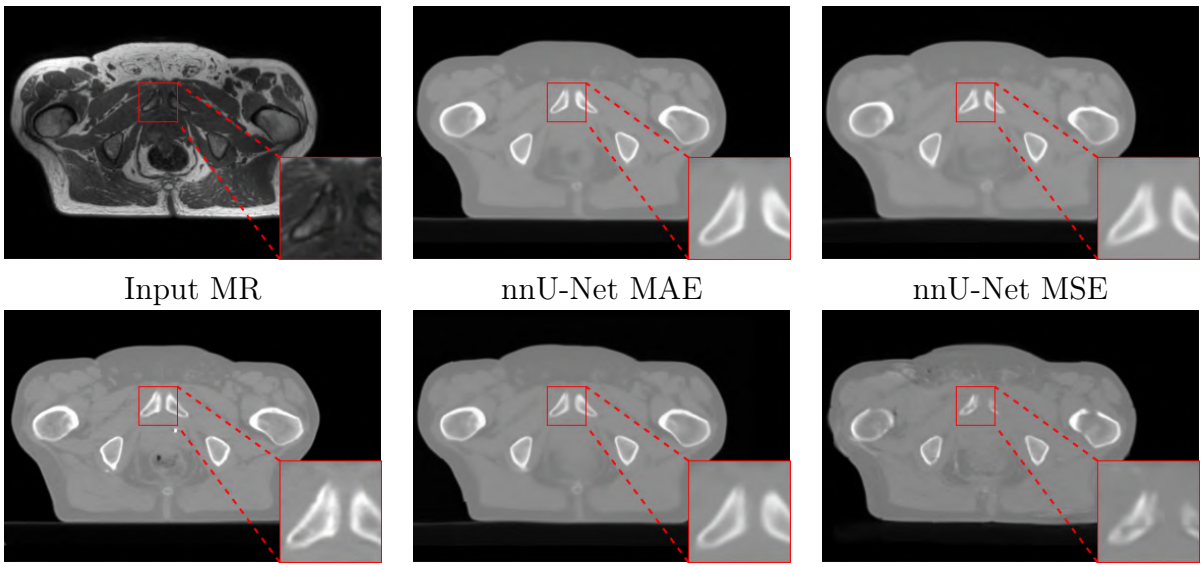


Figure 4.2: Comparison of axial slices between input MR image, ground truth CT, and synthesized CT from several nnU-Net implementations and 3D pix2pixHD model. Axial slices are presented, with zoomed-in regions.

In Figure 4.2, a comparison of MR to CT synthesis on the validation set is shown. We

observe that nnU-Net with MAE or MSE both perform exceptionally well in reconstructing CT scans, demonstrating greater precision than pix2pixHD, especially for bone reconstruction and body contouring. nnU-Net with MedicalNet also shows visually high-quality results, challenging the accuracy of the intensity metrics reported in Table 4.1, which indicated mitigated performance for this model. This discrepancy suggests that visual assessment may provide additional insights beyond the standard intensity-based metrics.

Brain Local Inpainting

In addition to evaluating MR to CT translation, we also applied our adapted nnU-Net to the task of brain image inpainting to assess how a highly optimized CNN architecture compares to leading GAN and diffusion model approaches. This task involves reconstructing voided healthy regions in T1-weighted brain MRIs, a problem typically tackled by models designed to restore missing or damaged parts of medical images.

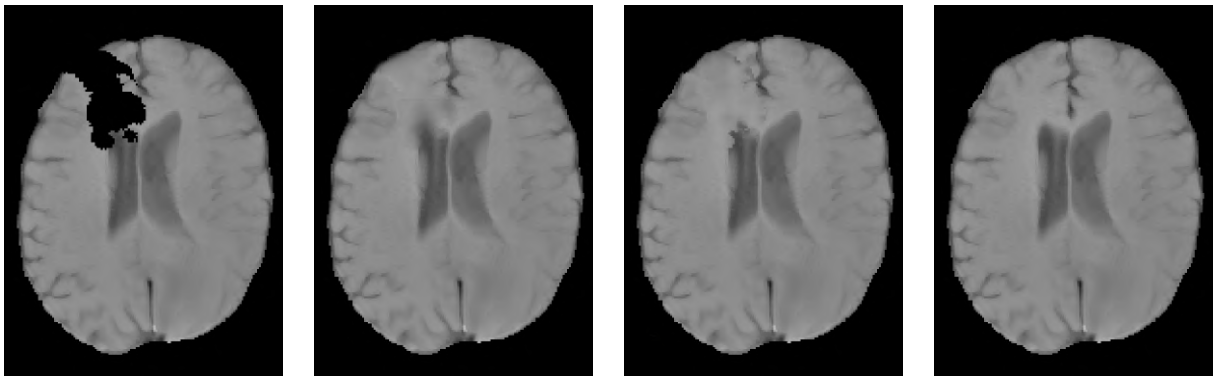
Table 4.2: Comparison of 3D networks performances using MSE, SSIM, and PSNR metrics between inpainted and original brain images. Evaluation is performed on 238 patients.

Model	MSE	SSIM	PSNR
3D nnU-Net MAE	0.0129 ± 0.0063	$\mathbf{0.797 \pm 0.113}$	19.45 ± 2.63
3D nnU-Net MSE	$\mathbf{0.0116 \pm 0.0056}$	0.797 ± 0.115	$\mathbf{19.87 \pm 2.55}$
3D nnU-Net MedicalNet	0.0142 ± 0.0064	0.749 ± 0.143	18.86 ± 2.11
3D pix2pixHD	0.0141 ± 0.0078	0.787 ± 0.129	19.28 ± 3.11
Results from available participants reports			
N/A - J. Zhang et al. [142]			
3D pix2pix	$\mathbf{0.0111 \pm 0.0068}$	$\mathbf{0.807 \pm 0.129}$	$\mathbf{19.30 \pm 2.49}$
2nd - A. Durer et al. [138]			
DDPM 2D slice-wise	0.0160 ± 0.0118	0.785 ± 0.155	18.71 ± 3.08

The quantitative validation results for the local inpainting task are presented in Table 4.2. Additionally, results from participant reports on the BraTS local inpainting test dataset are included, with the key difference being that the evaluations were conducted on the test set, whereas our approach utilized a subpart of the training set due to the closure of the BraTS challenge validation server. This table shows that our adapted nnU-Net with a standard MSE loss outpeforms both 3D pix2pixHD and 2D DDPM, and closely matches the 3D pix2pix.
Figure 4.3 presents an example of a T1-weighted input slice alongside its corresponding ground truth, and reconstruction from best nnU-Net and 3D pix2pix models. Both models attempt to reconstruct the voided areas but differ in their approaches. The nnU-Net, which optimizes the MSE loss known to smooth output intensities, tends to produce blurred regions while preserving general image structure. Conversely, the 3D pix2pixHD model, which also optimizes the GAN loss, attempts to recreate these structures but results in outputs that appear unnatural and show distinct boundaries with the surrounding image areas.

Computational resources

Table 4.3 provides a comparison of the computational resources required by different network models in terms of GPU memory usage, number of parameters, and training time



T1w voided | nnU-Net MSE | 3D pix2pixHD | Ground-Truth

Figure 4.3: Comparison of axial slices between the T1w voided image, reconstructions from the best nnU-Net and 3D pix2pixHD models, and the original T1w image

Table 4.3: Comparison of networks computational resources in terms of memory cost on GPU, number of parameters and training time per iteration. Here, (G) denotes the generator and (D) the discriminator.

Model	Memory Cost	Parameters	Time per iter.
nnU-Net MAE	8.3 GB	30.8 M	0.150 s
3D Pïx2PixHD	15.9 GB	547.3M (G) + 22.1M (D)	0.630 s

per iteration. The resources are evaluated for both the generator (G) and the discriminator (D) components where applicable. This table highlights the differences in computational efficiency and resource demands among the models used for MR to CT translation and local inpainting tasks.

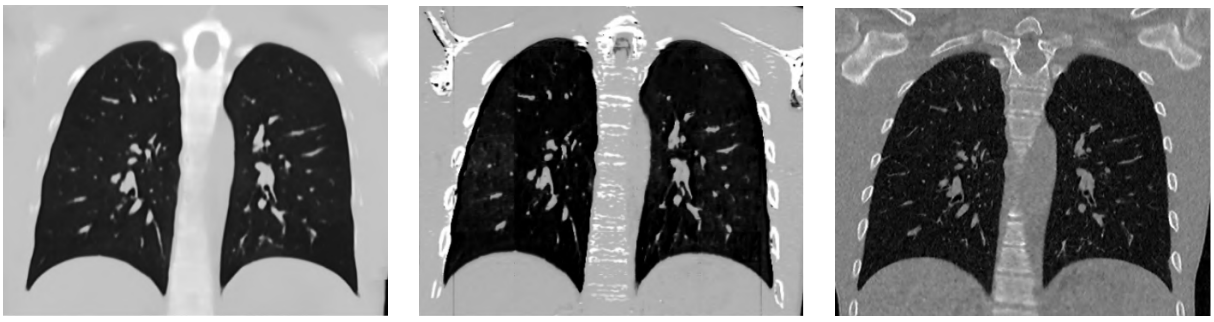
4.1.4 Findings and Limitations in Thoracic Imaging

This study presents a robust baseline for medical image synthesis based on an adapted nnU-Net framework. Our findings reveal that this method effectively reconstructs medical images and consistently outperforms GAN-based approaches, challenging the prevalent reliance on such methods within the field. Specifically, for the two evaluated tasks, Pelvis MR to CT translation and brain local inpainting with healthy reconstruction, the adapted nnU-Net, configured with a simple Mean Absolute Error (MAE) loss, produces solid results without the need for a discriminator.

Reproducibility and Consistency of Results The success of the adapted nnU-Net can be attributed to two key aspects of the framework. Firstly, nnU-Net's inherent adaptability and optimization capabilities ensure that it is well-suited to a variety of medical imaging tasks. It is designed to automatically configure itself optimally for the task at hand, making it highly versatile and effective. Secondly, the use of a U-Net architecture allows the model to focus directly on the defined cost function, in contrast to traditional GANs, which rely on discriminators that prioritize generating images that visually resemble the training dataset. Furthermore, the use of a simple U-Net architecture results in a lighter, more efficient model. As detailed in Table 4.3, nnU-Net operates with fewer parameters and less computational costs, while also reducing training times compared to the more complex pix2pixHD. This efficiency not only facilitates faster training cycles and

inference but also reduces the likelihood of overfitting or convergence issues commonly associated with GANs.
Another crucial aspect of this adapted nnU-Net is its contribution to the reproducibility of results. Reproducibility is often a challenge in medical image synthesis, where slight variations in model implementation can lead to significantly different outcomes. The nnU-Net framework helps standardize preprocessing and model configuration, which, as observed in the various implementations of models like pix2pix during the SynthRad2023 challenge [42], can lead to disparate results. By automating many aspects of its setup, nnU-Net makes it easier to replicate results, ensuring that outcomes are consistent and reliable across different studies.
For the inpainting task, while nnU-Net delivers robust performance in terms of intensity-based metrics, it tends to produce images with blurry tissues due to the use of MAE/MSE loss. Although effective in minimizing intensity errors, these losses fail to ensure detailed anatomical structure preservation, highlighting the need for more suitable loss functions. Additionally, recent works on pseudo 3D diffusion models [138] indicate they surpass other methods in the inpainting task, suggesting that diffusion models are particularly well-suited for these challenges. These models often incorporate a U-Net for the denoising process, suggesting that our adapted nnU-Net could serve effectively as this backbone.

Limitations in Thoracic Imaging While the nnU-Net framework demonstrates robust performance across a variety of tasks, some limitations became apparent when applying it to thoracic imaging, particularly in our initial task of lung MR to CT synthesis. The main challenge in thoracic imaging is the difference in spatial resolution between MR and CT modalities. The MR images tend to have lower resolution and often contain less detailed anatomical information than CT, making it difficult to achieve sharp reconstructions using traditional L1 or MSE loss functions. This often results in blurred outputs, as seen in our earlier lung MR to CT experiments, where finer details like small airways and vessels are not fully preserved.



nnU-Net MAE | nnU-Net MedicalNet | Real CT

Figure 4.4: Example of synthesized lung CT in coronal view using the adapted nnU-Net with L1 (MAE) loss and MedicalNet perceptual loss. The nnU-Net with L1 loss appears blurry and fails to reconstruct the region outside the lungs compared to real CT, while the perceptual loss with 3D MedicalNet embeddings is sharper but presents some artifacts and lacks precision in the reconstruction of fine structures.

To address this, we explored the use of 3D perceptual loss, employing MedicalNet embeddings, which provides sharper outputs but contain some artifacts on anatomical structure.

Figure 4.4 shows an example of synthesized lung CT using nnUNet with L1 and perceptual loss. One key issue is that MedicalNet was trained on a relatively small and less diverse dataset compared to larger datasets like ImageNet, which are used in traditional perceptual loss functions for natural images. This lack of diversity may limit the model's ability to generalize across different anatomical regions, resulting in artifacts and inaccuracies in synthesized images, particularly in complex thoracic structures like airways and vessels.

Need for Task-Specific Loss Functions In medical image synthesis, we primarily rely on global intensity-based loss functions, which often fall short in capturing the fine anatomical details required for specific tasks. While recent efforts have introduced 3D perceptual losses, such as those leveraging MedicalNet embeddings, these approaches can also struggle in complex scenarios due to the limited scale and diversity of the MedicalNet dataset.
To address these challenges, particularly in thoracic imaging, we plan to enhance the nnU-Net framework by developing task-specific loss functions that prioritize preserving anatomical structures over global intensities. These losses will be designed to capture the fine details required for accurate reconstructions of critical structures, such as bronchi and vessels, thereby improving the clinical utility of synthesized images.

4.2 Anatomical Feature-Prioritized Loss for Enhanced Synthesis

As demonstrated in the previous section, the adaptation of nnU-Net for 3D medical image synthesis has yielded promising results, particularly when compared to GAN-based approaches. However, while these methods effectively capture global image intensity and general structure, they often fall short when it comes to preserving fine anatomical details crucial for clinical applications. This is especially problematic in lung cross-modality synthesis tasks, where critical structures such as bronchi in lung imaging or organ boundaries in pelvic imaging must be accurately reconstructed for diagnostic or treatment purposes. To address these limitations, this section introduces an advanced loss function tailored for medical image synthesis: Anatomical Feature-Prioritized (AFP) loss. The goal of AFP loss is to prioritize the accurate reconstruction of anatomical structures, ensuring both global fidelity and fine detail preservation. Building on existing approaches that incorporate anatomical constraints, such as segmentation-based losses, we extend this idea by leveraging feature representations from models pre-trained on medical tasks, allowing for a more nuanced understanding of anatomical structures throughout the synthesis process. In this section, we detail the development of AFP loss, its integration into various generative models, and its evaluation on lung and pelvis MR to CT translation tasks.

4.2.1 Introduction

Over the past decade, medical image synthesis has emerged as a significant trend in the field of healthcare [143]. This technique plays a pivotal role in producing synthetic data to address the shortfall in medical datasets, such as data diversity [144], and improving the robustness of machine learning models. Additionally, it serves as a valuable tool for

translating between different imaging modalities, offering benefits in applications like attenuation correction or radiation therapy planning, as described in [62]. The significant advancement of this field is mainly attributed to the evolution of deep learning, notably the advent of generative methods, such as generative adversarial networks (GANs) [25] and diffusion models (DMs) [32]. Many recent studies have explored synthetic data generation [145] as well as cross-modality translation, addressing reconstruction challenges in anatomical areas such as brain [48], aorta [64], lungs [135].

Despite these advances, there remains a gap in the precision and fidelity of synthesized images, particularly in the reconstruction of detailed anatomical structures. Most current state-of-the-art methods, including those based on conditional generative models (cGANs, introduced in [36]) focus on optimizing global image reconstruction, often at the expense of accurately capturing the nuances of specific features. This is largely due to the reliance on global loss functions such as the L1 loss, which primarily measures global intensity differences and does not account for the preservation of anatomical structures and critical medical features. This kind of loss also tends to smooth out finer details, leading to images that may appear realistic at a macro level but lack the precision required for clinical tasks involving the identification and precise measurement of small, intricate structures. This oversight becomes particularly critical when addressing explicit clinical needs, such as the accurate delineation of boundaries between different tissues, the identification of lesions, or the precise measurement of anatomical structures for surgical planning or disease monitoring.

Several papers have focused on cross-modality synthesis based on anatomical structures. In [146], authors use a structure-constrained GAN for unsupervised MR to CT synthesis by defining a structure-consistency loss based on the modality independent neighborhood descriptor, introduced in [75], often employed for cross-modality registration. [147] employs a multi-task network for synthetic CT generation from MRI, emphasizing bone density value prediction. It uses a composite loss function to localize regions of interest, combining classification and regression tasks to improve performance and computational efficiency, particularly in radiation therapy planning. [148] introduces a structure-aware GAN for organ-preserving synthetic CT generation from MRI, using a dual-stream approach based on a segmentation loss to maintain organ integrity and achieve clinically acceptable accuracy in MR-only treatment planning.

These methods focus on specific body parts or anatomical structures using segmentation networks, but they do not use the full capabilities of these networks. While employing a segmentation loss is interesting for cross-modality synthesis, it typically relies only on the activation of the last layer of the network and is limited by the specific labels used, neglecting the rich representations available across various network layers. Addressing these limitations, our study introduces a novel approach aimed at enhancing the synthesis of specific medical details across different anatomical regions. We introduce an anatomical feature-prioritized loss function, denoted AFP loss, specifically designed for medical image translation, ensuring both global fidelity and detailed structural accuracy. This approach leverages a distance between features extracted from a model pre-trained on a specific medical task, such as the segmentation of a delimited region, based on the idea that such a model presents an advanced understanding and representation of those medical structures.

We integrate this feature-prioritized loss function into various synthesis networks, including generative models like GANs (pix2pixHD from [43], SPADE from [44]) and adapted CNN models (nnU-Net from [93]), to evaluate their effectiveness comprehensively. We

then assess the performance of our approach on two distinct tasks: MR to CT synthesis in the lungs, aiming for fine structure generation including vessels and bronchi, and MR to CT synthesis in the pelvis, focusing on the generation of coarser anatomical features such as organs and muscles. Our study provides a robust qualitative and quantitative evaluation, focusing on the precise reconstruction of anatomical structures in these regions, demonstrating the potential of our approach to improve clinical outcomes.

4.2.2 Definition

In this study, we introduce a novel anatomical feature-prioritized loss function (denoted AFP loss). The primary goal of this loss is to enhance the synthesis of specific anatomical structures in medical images, which are often inadequately reconstructed when relying solely on standard reconstruction losses. Our approach leverages a pre-trained segmentation convolutional network, specifically tailored to segment precise anatomical regions. The specialization of this network in segmenting precise regions allows us to leverage its data representations to improve the generation of anatomical features. The mechanism of our feature-prioritized loss mirrors that of a perceptual loss described in [106], primarily focusing on minimizing the difference between the feature representations of synthetic and real images extracted from various layers of a CNN. Traditionally, perceptual loss is defined using 2D classification networks on natural images, such as VGG on ImageNet [108], and its use in 3D is often limited. Our anatomical feature-prioritized loss, in contrast, leverages segmentation networks, making it more straightforward to implement in medical tasks. Additionally, it benefits from more precise feature representations compared to those in classification models. The AFP loss is formulated as follows :

$$\mathcal{L}_{AFP(x,y)} = \frac{1}{N}\sum_{i=1}^{N} \|\phi_i(x) - \phi_i(y)\|_1 \tag{4.1}$$

Here, x and y represent the synthesized and real images to compare, N denotes the number of layers extracted from the pre-trained network and ϕ_i the features of the i-th layer. Figure 4.5 illustrates the AFP loss based on the embeddings of an airway segmentation model.

Integration in synthesis networks

The AFP loss introduced in this study is a versatile function, adaptable to various synthesis and translation tasks within medical imaging. Depending on the specific synthesis task, the AFP loss can be combined with a global reconstruction loss, such as the traditional L1 loss, to maintain a correct range of intensities. A general recommendation is to start with a global network trained with L1 loss to establish a baseline intensity reconstruction and then fine-tune the model with our AFP loss to refine the anatomical structures. This approach ensures a two-step process: an initial global generation step followed by a local anatomical refinement step.
For tasks involving the generation of fine structures, such as vessels or bronchi, it may be preferable to use only the AFP loss from the beginning. In such cases, L1 loss should be avoided due to its tendency to produce blurred reconstructions, which can obscure fine details. The AFP loss excels in these scenarios by focusing on precise anatomical fidelity without being influenced by global intensity differences.
While we illustrate this loss using pre-trained U-Net models, our approach can be adapted

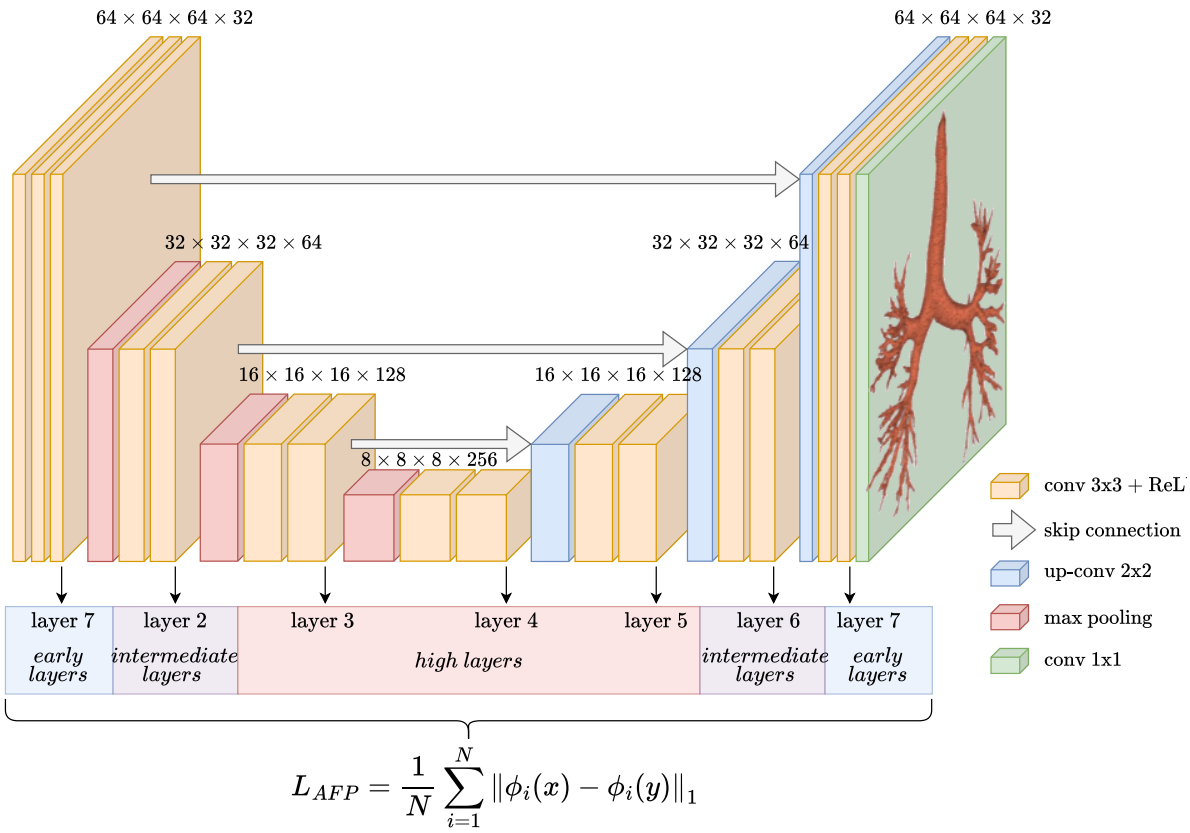


Figure 4.5: Illustration of the Anatomical Feature-Prioritized (AFP) loss on a U-Net network trained for airways segmentation. All features are acquired after the ReLU activation in each convolutional block. Layer 7 is extracted before the final convolution.

to any CNN trained on medical data. The AFP loss serves as a robust addition to medical image synthesis and translation tasks, enhancing the capability to generate high-fidelity anatomical structures.

Implemented Synthesis Models

To evaluate the performances of our anatomical structure prioritized loss function (AFP loss), we implemented it in two primary models for medical image synthesis: a GAN-based model and a CNN-based models. Specifically, we develop a 3D variant of the pix2pixHD model and adapt a 3D nnU-Net model for translation tasks.

GAN-based model We develop a 3D variant of the pix2pixHD model [43], incorporating 3D versions of the generator and discriminator, as well as 3D SPADE layers [44]. This architecture facilitates image-to-image translation by incorporating spatially-adaptive normalization, effectively mitigating the issue of false positives encountered in the pix2pixHD network alone [109]. For the training process, we employ a composite loss function that includes an adversarial loss, a feature-matching loss, and a global reconstruction loss. The adversarial loss utilizes a Hinge loss mechanism, as adopted in the SPADE implementation [44], to refine the adversarial training dynamics. The feature-matching loss ensures alignment between the discriminator's intermediate representations of the real and synthesized images, utilizing an L1 distance to promote stability and fidelity in

the generated images [43]. A global reconstruction loss is incorporated to further enhance the overall quality of the synthesized output, defined as an L1 loss. Finally, our AFP loss is incorporated to complement the L1 loss, enhancing localized structural details, or to replace it entirely, offering a more targeted approach to preserving anatomical accuracy in the synthesized images.

CNN-based model We have extended the application of the nnU-Net framework [93], traditionally a pioneer in medical image segmentation, to tackle translation tasks. Our previous work [149], presented in Section 4.1 provides an adaptation of the nnU-Net for synthesis tasks and showed promise with a standard MSE loss function. The adapted model retains the original framework's preprocessing and low-complexity architecture, which benefits from not being dependent on a discriminator, a component known for its training challenges. We propose a new version of the adapted nnU-Net incorporating the proposed AFP loss in the training process. This setup enables us to precisely quantify the impact of our AFP loss, whether used alone or paired with an L1 loss, providing a robust framework for improving medical image synthesis. The model has been made publicly available at https://github.com/Phyrise/nnUNet_translation.

Reconstruction from patches

To address the memory constraints associated with processing 3D models, we adopt a patch-based approach for each synthesis model. To ensure smooth reconstruction and mitigate border effects between patches, we utilize standard median reconstruction on voxel intensities, with patches defined by a tiling of 0.5. This method preserves high-frequency details, resulting in a visually more accurate and detailed final output. While mean reconstruction often acts as a smoothing filter that can perform better on certain metrics, median reconstruction provides superior visual quality by maintaining the finer details.

Checkerboard artifacts

The AFP loss operates similarly to a perceptual loss, which has been shown to potentially lead to checkerboard artifacts [139], primarly due to transposed convolution operations in the decoder part. To adress this issue, we explore the effectiveness of combining the AFP loss with a traditional L1 loss, which can be effective in reducing high-frequency errors that contribute to checkerboard patterns. This approach aim to leverage the global intensity preservation of the L1 loss while benefiting from the detailed structure preservation of the AFP loss. We also investigated replacing transposed convolutions with a combination of upsampling and convolutional layers as proposed by [140]. This proven technique helps to reduce checkerboard artifacts by providing a more stable upsampling mechanism, particularly when using feature-based losses.

4.2.3 Validation on MR to CT translation

Datasets

Lung MR to CT Synthesis For lung MR to CT synthesis, we used a private dataset of thoracic images from UTE MR and CT scans of 122 patients. Both modalities were acquired on the same day, from 2018 to 2022. CT images were obtained using a Siemens

SOMATOM Force and a GE Medical Systems Revolution CT in end-expiration, with sharp filters. The acquisition parameters were a DLP of 10 mGy.cm and a SAFIRE iterative reconstruction. The UTE MR images were acquired using the SpiralVibe sequence on a SIEMENS Aera scanner, with the following parameters: TR/TE/flip angle=4.1ms/0.07ms/5°. Since the slice plane is encoded in Cartesian mode, the native acquisition was performed in the coronal plane with field-of-view outside the anterior and posterior chest edges to prevent aliasing. It should be noted that resolutions, voxel spacings, and fields of view are not identical in CT and MR images. In addition, modalities may have been taken at different points in the respiratory cycle.
All volumes are resampled to have a voxel size of $0.6 \times 0.6 \times 0.6$ mm^3. To align CT images with MR images, we first perform a rigid translation, followed by a deformable registration process using the EVolution algorithm proposed by [76]. This method incorporates a similarity term that prioritizes the alignment of edges and employs a diffeomorphic transformation to preserve the CT volume topology. Subsequently, we normalize MR images to have zero mean and unit variance, and for CT images, we use intensity values from the foreground classes to compute the mean and standard deviation, clip values to the 0.5 and 99.5 percentiles, and then normalize by subtracting the mean and dividing by the standard deviation.

Pelvis MR to CT Synthesis For the synthesis of pelvis images, we used the open-access dataset from Task 1 of the SynthRAD2023 challenge from [41]. This dataset includes paired MR and CT images from 180 patients used for training, with an additional 30 patients designated for validation. The dataset provides high-quality, well-aligned MR and CT image pairs, which eliminates the need for additional registration processes. The voxel sizes in this dataset are already standardized to $2.5 \times 1 \times 1$ mm, ensuring consistent spatial resolution across all images. Similar preprocessing steps from the lung dataset are applied, with z-score normalization for MR images and clipping CT values to the 0.5 and 99.5 percentiles, followed by normalization using the mean and standard deviation.

Evaluation Metrics

To comprehensively assess our synthesis models, we employ a dual approach combining intensity-based and task-specific metrics.

Intensity-based Metrics Traditional intensity-based metrics, such as Mean Absolute Error (MAE) and Structural Similarity Index (SSIM), provide a measure of global image synthesis quality but may not adequately capture the nuances required for specific medical tasks. Accurate delineation of boundaries between different tissues, identification of lesions, and precise measurement of anatomical structures for surgical planning or disease monitoring necessitate evaluation metrics based on the reconstruction of anatomical structures. Although qualitative assessments by expert radiologists are valuable, they are also very tedious and time-consuming.

Task-specific Metrics To capture the nuances required for specific medical tasks, we introduce a downstream task, the segmentation of anatomical structures, as a key evaluation method. This approach provides a more precise measure of the model's performance at the anatomical level and helps quantify the false positive rate, crucial for clinical applications. We use segmentation-based metrics like the Dice score and the Normalized Sur-

face Distance (NSD) [104] to evaluate local reconstructions. NSD is particularly valuable as it includes a manually defined margin of error, enhancing its utility in analyzing fine structures and providing robustness against registration inaccuracies [105]. In our study, the margin of error for NSD calculations is set at twice the voxel size: 1.2mm for lungs and 2mm for pelvis. To mitigate potential biases introduced by relying on pre-trained segmentation networks, we evaluate our synthesis models across multiple segmentation tasks and models. This ensures a comprehensive and balanced assessment of their performance. For both tasks, the segmentation of the real CT is defined as a silver-standard, given that the actual segmentation masks are not available.

Lung MR to CT Synthesis For the lung region, we use segmentations from three pre-trained models:

1. The public NaviAirway network [102], focused on segmenting airways and bronchi.
2. The private Holistic Airway Lesions (HAL) [150] model measures anatomical structures within the lungs, such as bronchiectasis and peribronchial thickening.
3. The publicly available TotalSegmentator [151] models, which have been fine-tuned and fused into one comprehensive model. This network focuses on coarser segmentations such as lung lobes, bones, muscles, and organs outside the lungs.

Pelvis MR to CT Synthesis For the pelvis region, we use segmentations from the TotalSegmentator model. This network provides segmentation of various tissues in the pelvic region, including bones (e.g., sacrum, hips), organs (e.g., bladder, prostate), and muscles (e.g., iliopsoas, gluteus).
This dual approach of intensity-based and task-specific metrics allows for a comprehensive evaluation of both global image quality and specific anatomical structure preservation in our synthesis models.

Results

Lung MR to CT Synthesis In the lung MR to CT synthesis task, we established the baseline using the traditional L1 loss and a perceptual loss leveraging MedicalNet's embeddings. We evaluated these losses against our proposed AFP loss, using various pre-trained segmentation networks: TotalSegmentator (TotalSeg), NaviAirway (Navi), and Holistic Airway Lesions (HAL). Additionally, a hybrid approach combining NaviAirway and Holistic Airway Lesions (denoted as Navi/HAL) was explored to integrate embeddings from both healthy and unhealthy patients within the airway region.
Figure 4.6 presents a visual comparison of lung coronal sections from different models. The model using L1 loss results in blurred images, with particularly poor bone reconstruction. In contrast, the MedicalNet-based model produces sharper images but introduces artifacts in anatomical areas. Models based on AFP loss offer visually appealing results, with well-reconstructed anatomical bones and lung tissue. Additionally, only NaviAirway and HAL methods provide detailed bronchial reconstruction, as highlighted in the zoomed-in section.
Table 4.4 presents a quantitative evaluation of the model's performance on lung MR to CT synthesis, based on the MAE, SSIM, Dice score, and NSD between synthesized and ground truth CT images. The adapted nnU-Net trained with L1 loss delivers the best

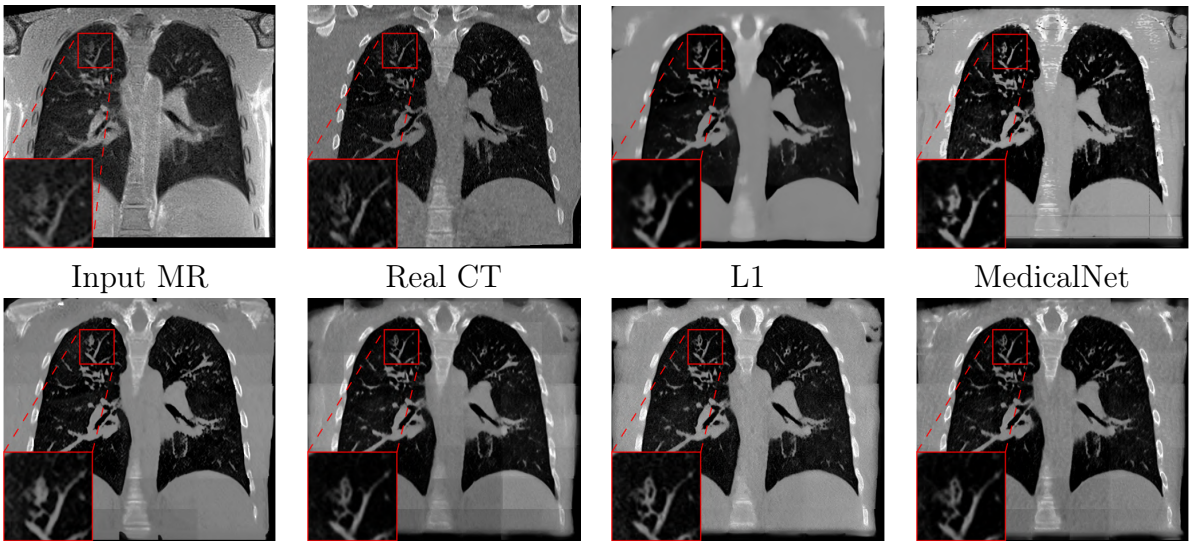


Figure 4.6: Comparison between input MR image, ground truth CT, and synthesized CT from adapted nnU-Net with several loss functions. Coronal slices are presented, including a zoomed-in view of the bronchi reconstruction. The zoomed-in view demonstrates that AFP loss with NaviAirway and HAL features enables the most faithful reconstruction of the real CT.

Table 4.4: Comparison of 3D lung networks performances using MAE and SSIM between real CT and synthesized CT, as well as Dice score and NSD on airways segmentation using the NaviAirway pipeline. Results include adapted nnU-Net model with various losses and GAN-based SPADE model.

	Intensity-based metrics		**Task-based metrics**	
Model / Loss	**MAE**	**SSIM**	**Dice**	**NSD**
nnU-Net L1	$\mathbf{48.72 \pm 12.84}$	$\mathbf{0.837 \pm 0.015}$	0.534 ± 0.066	0.644 ± 0.089
MedicalNet	69.66 ± 19.75	0.814 ± 0.016	0.471 ± 0.064	0.591 ± 0.074
AFP TotalSeg	59.40 ± 14.91	0.828 ± 0.014	0.530 ± 0.068	0.667 ± 0.076
AFP Navi	64.98 ± 16.88	0.824 ± 0.014	0.580 ± 0.059	0.720 ± 0.064
AFP HAL	60.92 ± 15.90	0.827 ± 0.014	0.564 ± 0.067	0.702 ± 0.079
AFP Navi/HAL	61.47 ± 15.98	0.827 ± 0.014	$\mathbf{0.584 \pm 0.063}$	$\mathbf{0.723 \pm 0.072}$
SPADE L1	55.68 ± 14.42	0.829 ± 0.015	0.521 ± 0.062	0.623 ± 0.085
SPADE AFP Navi	69.15 ± 18.29	0.815 ± 0.016	0.545 ± 0.071	0.689 ± 0.079



Figure 4.7: Example of 3D airway segmentation using the NaviAirway pipeline. The trachea, in transparency, is excluded for the calculation of metrics. The AFP loss facilitates the reconstruction of higher bronchi compared to the L1 loss, as depicted in Table 4.4.

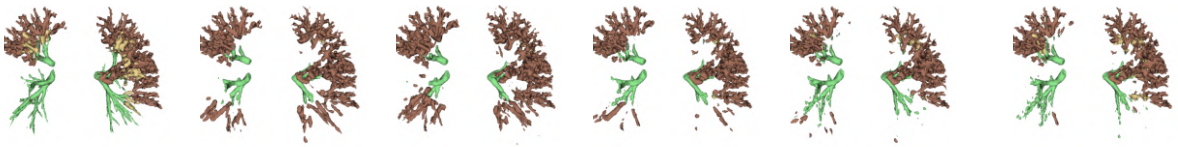
performance on intensity-based metrics. Other models based on perceptual loss or AFP loss achieve average results in terms of MAE but maintain reasonable SSIM values. In the context of airway segmentations using the NaviAirway pipeline, the adapted nnU-Net models with AFP loss from NaviAirway and HAL's embeddings deliver the best performance, with the highest Dice scores and NSD values. Conversely, models employing L1, perceptual, or AFP loss with TotalSegmentator's embeddings yield poorer results, lacking precise bronchial reconstruction. The GAN-based SPADE method generally underperforms compared to the adapted nnU-Net, but adding AFP loss to SPADE enhances its performance. These metrics align with qualitative analysis from Figure 4.6 and Figure 4.7, with the models using AFP loss delivering the best performance in airway reconstruction.

Table 4.5 and Figure 4.8 provide a comparative evaluation of the models on airway lesions segmentations using the Holistic Airway Lesions (HAL) pipeline with Dice score and NSD. Models leveraging the AFP loss using HAL's embeddings provide the best performances. Additionally, the model with AFP loss using only NaviAirway's embeddings also shows strong results, while other models are generally of poor quality on this task.

Table 4.6 and Figure 4.9 present a comparison of the models on the reconstruction of lung anatomical regions using the TotalSegmentator (TotalSeg) pipeline with Dice score and NSD. The model trained with the AFP loss using TotalSeg's embeddings provide the best performances. A combination of AFP loss using NaviAirway and HAL's embeddings also delivers positive results, whereas other models exhibit notably poorer reconstructions especially in the bones.

Table 4.5: Comparison of 3D nnU-Net models with different losses for lung synthesis, evaluated using Dice score and NSD on Airway Lesions segmentations between real CT and synthesized CT.

Loss	**Bronchiectasis**		**Bronchial Mucus**	
	Dice	**NSD**	**Dice**	**NSD**
L1	0.287 ± 0.087	0.373 ± 0.108	0.281 ± 0.094	0.369 ± 0.146
MedicalNet	0.237 ± 0.087	0.308 ± 0.110	0.225 ± 0.095	0.276 ± 0.132
AFP TotalSeg	0.308 ± 0.086	0.376 ± 0.119	0.303 ± 0.094	0.375 ± 0.146
AFP Navi	0.363 ± 0.102	0.500 ± 0.145	0.358 ± 0.101	0.447 ± 0.119
AFP HAL	$\mathbf{0.393 \pm 0.105}$	$\mathbf{0.516 \pm 0.127}$	$\mathbf{0.363 \pm 0.109}$	0.472 ± 0.125
AFP Navi + HAL	0.388 ± 0.096	0.496 ± 0.122	0.362 ± 0.099	$\mathbf{0.484 \pm 0.134}$



Real CT | L1 | MedicalNet | AFP TotalSeg | AFP Navi | **AFP Navi/HAL**

Figure 4.8: Comparison of airways lesions segmentations from real CT and synthesized CTs using the Holistic Airway Lesions (HAL) pipeline for an unhealthy patient. Bronchiectasis is represented in green, peribronchial thickenning in yellow and bronchial mucus in brown. Results correlate with Table 4.5, the AFP loss demonstrates the best visual outcomes, delivering fewer false positives in bronchial mucus compared to the L1 or MedicalNet losses.

Table 4.6: Comparison of 3D nnU-Net models with different losses for lung synthesis, evaluated using Dice score and NSD on TotalSegmentator segmentations between real CT and synthesized CT.

Loss	Muscle/Organs		Bones	
	Dice	NSD	Dice	NSD
L1	0.794 ± 0.053	0.871 ± 0.058	0.402 ± 0.081	0.493 ± 0.100
MedicalNet	0.774 ± 0.043	0.850 ± 0.048	0.405 ± 0.058	0.499 ± 0.071
AFP TotalSeg	$\mathbf{0.811 \pm 0.057}$	$\mathbf{0.900 \pm 0.063}$	$\mathbf{0.516 \pm 0.070}$	$\mathbf{0.629 \pm 0.086}$
AFP Navi	0.751 ± 0.063	0.801 ± 0.081	0.479 ± 0.074	0.583 ± 0.092
AFP HAL	0.785 ± 0.064	0.859 ± 0.071	0.481 ± 0.076	0.590 ± 0.092
AFP Navi/HAL	0.786 ± 0.060	0.844 ± 0.066	0.499 ± 0.074	0.604 ± 0.090



Real CT | L1 | MedicalNet | **AFP TotalSeg** | AFP Navi | AFP Navi/HAL

Figure 4.9: Comparison of lung labels segmentations from real CT and synthesized CTs using the TotalSegmentator pipeline. Bones are represented in yellow, lung lobes in blue, green, brown and red colors. Visual results align with Table 4.6 : The AFP loss achieves superior reconstruction, particularly in the bones.

Pelvis MR to CT Synthesis

Similarly to the lung task, in the pelvic image synthesis task, we established a baseline using the traditional L1 loss and perceptual loss using the MedicalNet model. Our AFP loss, leveraging embeddings from the TotalSegmentator (denoted TotalSeg) network, was evaluated against this baseline. Furthermore, we explored a hybrid approach that combines L1 loss with AFP loss (denoted L1 + AFP) to assess the benefits of retaining the global reconstruction L1 loss.

Table 4.7 presents a quantitative evaluation of the model's performance on pelvic MR to CT synthesis, based on the MAE and SSIM between synthesized and ground truth CT images. Similarly to the lung results, the model trained with L1 loss delivers the best performance on intensity-based metrics. Other models based on perceptual loss or AFP loss alone achieve average results in terms of MAE but maintain reasonable SSIM values. Finally, the combination of L1 and AFP losses provides solid results on these intensity-based metrics, bringing it closer to the high standards set by L1 loss.

Table 4.7: Comparison of 3D nnU-Net models with different losses for pelvic synthesis using MAE and SSIM between real CT and synthesized CT.

Loss	MAE	SSIM
L1	$\mathbf{65.300 \pm 15.180}$	$\mathbf{0.852 \pm 0.035}$
MedicalNet	81.359 ± 30.717	0.838 ± 0.038
L1 + AFP TotalSeg	66.937 ± 15.116	0.847 ± 0.035
AFP TotalSeg	85.258 ± 14.471	0.839 ± 0.036

Figure 4.10 presents a visual comparison of pelvic sagittal sections from different models.

The model trained with L1 loss displays blurry results, with smooth intensities across the body and a noticeable lack of precise structural delineations. Conversely, the MedicalNet-based model, while producing sharper images, tends to introduce artifacts and also falls short in defining precise anatomical structure borders. In contrast, the TotalSegmentator model delivers excellent visual quality, featuring sharp reconstruction and precise delineation of structures, notably in the colon region as depicted in the zoomed-in preview.

Table 4.8: Comparison of 3D nnU-Net models with different losses for pelvic synthesis using Dice score and NSD on TotalSegmentator segmentations between real CT and synthesized CT.

Loss	**Bones**		**Muscles**	
	Dice	**NSD**	**Dice**	**NSD**
L1	0.738 ± 0.239	0.774 ± 0.222	0.687 ± 0.210	0.719 ± 0.208
MedicalNet	0.727 ± 0.241	0.759 ± 0.236	0.646 ± 0.237	0.698 ± 0.225
AFP TotalSeg	$\mathbf{0.780 \pm 0.203}$	$\mathbf{0.823 \pm 0.200}$	0.724 ± 0.193	0.762 ± 0.188
L1 + AFP	0.772 ± 0.194	0.817 ± 0.196	$\mathbf{0.728 \pm 0.195}$	$\mathbf{0.768 \pm 0.185}$

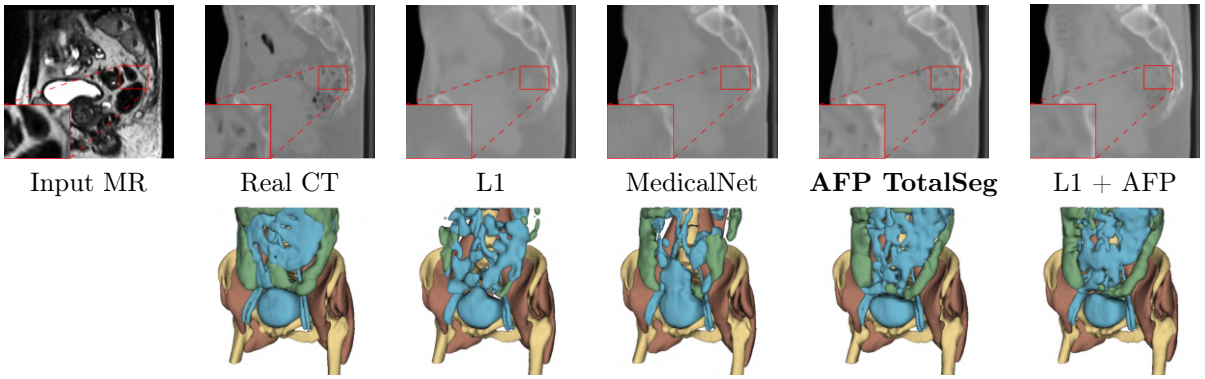


Figure 4.10: Comparison between input MR image, ground truth CT, and synthesized CT from nnU-Net translation with several loss functions. Sagittal slices are presented, including a zoomed-in view of the colon reconstruction on the top row, and 3D view of TotalSegmentator segmentations on the bottom row. Bones are represented in yellow, muscle in brown, organs in blue and colon region in green. Visual results align with Table 4.8 and show that AFP loss enables a better anatomical reconstruction, especially in the organs and colon.

Table 4.8 and bottom row of Figure 4.10 provide a comparative evaluation of the models on anatomical regions using the TotalSegmentator pipeline with Dice score and NSD. Aligning with the visual analysis, the model trained with the AFP loss using TotalSeg's embeddings provide the best performances in every region. Combining the AFP loss with a L1 loss does not significantly affect the reconstruction performance of anatomical structures. The model trained with MedicalNet perceptual loss provides poor results overall, while the model trained with L1 loss delivers average performance in each region.

4.2.4 Synthesis of Findings

Feature-Prioritized Loss: A Simple Yet Effective Approach This study presents an exploration into medical image synthesis with a focus on MR to CT translation for

both lung and pelvic regions, specifically aiming to enhance localized structural detail reconstruction through a novel anatomical feature-prioritized loss (AFP loss). Our findings reveal that this method effectively reconstructs targeted anatomical features and can either complement or replace global reconstruction strategies such as L1 loss or perceptual loss. The AFP loss, while conceptually simple in its implementation—essentially applying L1 loss in the embeddings of a pre-trained segmentation network—demonstrates a significant impact on the representation of anatomical structures. This straightforward approach allows the synthesis model to develop a more advanced understanding and representation of complex medical structures, highlighting the potential of leveraging domain-specific knowledge in synthesis networks.

Comparison with Traditional Methods and Metrics Global reconstruction loss is commonly expressed using a perceptual loss, an L1 loss, or an L2 loss. The standard L1 loss, directly applied to voxel intensities, is usually defined as the primary global reconstruction loss, as it is more robust to outliers in contrast to the L2 loss, which also tends to yield blurry results [39]. On the other hand, perceptual losses are often more robust and deliver better results than L1 loss but are usually limited to 2D applications [152], unlike our AFP loss, which can easily be used in 3D using trained segmentation networks. However, the effectiveness of the feature-prioritized loss is contingent upon the performance of the pre-trained model on the given task, as it relies on the auxiliary model's ability to provide meaningful gradients for the synthesis model's training. Our experiments showed that the AFP loss could significantly enhance the reconstruction quality of specific anatomical structures when integrated into various synthesis networks, including GAN-based models (pix2pixHD, SPADE) and CNN-based models (adapted nnU-Net).
For global image synthesis quality, traditional metrics such as MAE and SSIM were employed. However, these intensity-based metrics alone often do not adequately capture the nuanced needs of medical imaging, such as the accurate delineation of boundaries between different tissues or the identification of lesions. This underscores the necessity of using structure-specific metrics such as Dice score and Normalized Surface Distance (NSD) for evaluating the quality of local reconstructions, particularly focusing on preserving anatomical structural details that are vital for clinical analysis.

Performance and Insights Our experiments indicate that the AFP loss excels in refining the reconstruction of targeted structures, regardless of the dataset (lung or pelvis) or the method used (GAN or U-Net). In the lung region, we evaluated our models on three segmentation tasks: airways and sain bronchi using the NaviAirway pipeline, airway lesions with the Holistic Airway Lesions (HAL) pipeline, and larger anatomical regions (muscles, organs, bones) using the TotalSegmentator pipeline. As expected, the best model for each task is often the one with embeddings closely aligned with the task being evaluated. For instance, for muscle and organ reconstruction based on TotalSegmentator segmentation, the most effective model utilizes AFP loss with TotalSegmentator's embeddings.
However, surprisingly, some networks with non-specialized embeddings can produce high-quality reconstructions or even outperform specialized models when combined with them. For example, the combination of NaviAirway and HAL's embeddings achieved the best performance for airway reconstruction, surpassing the AFP loss with NaviAirway's embeddings alone. Additionally, for bone reconstruction, AFP loss combining NaviAirway and HAL's embeddings yielded excellent results, despite these two networks originally

being specialized in fine structures within the lungs, which are quite different from bone segmentation. These findings highlight the advanced understanding and representation capabilities of pre-trained segmentation networks and underscore the value of using them as replacements for perceptual loss or L1 loss.
Moreover, while a general model like TotalSegmentator is not perfectly suited for fine structure reconstruction, it serves as an excellent starting point for coarser tasks. It systematically outperforms L1 loss and MedicalNet perceptual loss in every segmentation task, and provides sharper and more realistic visual reconstructions compared to these other models.
In the pelvic region, models using AFP loss on TotalSegmentator's embeddings and AFP loss combined with L1 loss performed similarly in the downstream task and had similar visual reconstructions, despite having very different results in the intensity-based evaluation. This discrepancy once again calls into question the relevance of these metrics, as they do not accurately reflect the medical and anatomical accuracy of the synthesized volumes.
Initially, perceptual loss, typically derived from pre-trained 2D networks like VGG, was considered. However, when applied using a 3D MedicalNet model specifically adapted for medical data, the results were underwhelming, even when compared to the traditional L1 loss. This suggests ongoing challenges in using medical image datasets for perceptual loss due to their limited diversity and smaller scale compared to extensive, varied datasets like ImageNet [153]. Consequently, we recommend relying on the L1 loss or AFP loss with TotalSegmentator's embeddings for coarse reconstructions, which consistently delivered superior image quality.
Moreover, the elimination of the discriminator, particularly when using nnU-Net as a generator only, resulted in even better outcomes. This highlights the potential need to reconsider traditional cost functions in medical synthesis. Instead of focusing on output likelihood — which is typical with discriminators — a more clinically relevant approach would focus on preserving anatomical structures and ensuring low false-positive rates. This aligns directly with the clinical objective of synthesizing volumes that not only possess high signal quality but also maintain the integrity of anatomical details, facilitating advanced analysis by radiologists.

Chapter Conclusion

In this chapter, we expanded our focus beyond lung MR to CT synthesis to develop a more generalizable method for various medical image synthesis tasks. Using the nnU-Net framework, we created a scalable solution capable of addressing different synthesis challenges across multiple anatomical regions. To validate this approach, we applied it to public datasets for tasks such as pelvis MR to CT translation and brain inpainting. By adapting the nnU-Net architecture for 3D medical image synthesis, we established a reproducible and consistent baseline, outperforming GAN-based approaches, which often require complex adversarial training. With its automated configuration and task-specific optimization, nnU-Net proved to be more flexible, lighter, and efficient than diffusion models, making it a practical solution for the synthesis of medical images across diverse tasks and anatomical regions.

However, we identified limitations in using standard loss functions, such as L1 and 3D perceptual loss, which struggled to capture fine anatomical details. This motivated the development of our Anatomical Feature-Prioritized (AFP) loss, designed to focus on reconstructing critical structures. AFP loss represents a significant leap forward by integrating domain-specific knowledge into the synthesis process. By leveraging the feature embeddings of pre-trained segmentation networks, AFP loss allows for a focused reconstruction of essential anatomical features such as airways, organs, muscles and bones. This shift in priority improved intra-pulmonary synthesis tasks and demonstrated effectiveness in extra-pulmonary areas.

In conclusion, this chapter presents a comprehensive approach to medical image synthesis that combines the flexibility of nnU-Net with the anatomical precision of AFP loss. Our work fundamentally enhances the accuracy and reproducibility of synthesized medical images, offering a scalable solution for various medical imaging tasks. The validation of our methods on public datasets underscores their generalizability and clinical potential. Moving forward, our focus will be on extending AFP loss to other anatomical regions and refining its integration with dynamic weighting strategies for feature maps. By further enhancing the model's ability to balance global and local reconstructions, we aim to bring medical image synthesis closer to clinical adoption.

Conclusion and Perspectives

Summary of Findings

In our first study, we investigated the feasibility of using 2D generative adversarial networks (GANs) for MR to CT translation in the lungs. Given the complexity of lung anatomy and the challenges posed by respiratory motion, we optimized preprocessing steps to enhance the quality of the input data. This included refining image registration techniques to align MRI and CT images accurately, adjusting resampling strategies to achieve optimal spatial resolution, and standardizing intensity normalization and fields of view. To mitigate the common issue of GANs generating hallucinated features, a critical concern in medical imaging, we incorporated the Spatially-Adaptive (SPADE) normalization into our GAN architecture. The enhanced preprocessing pipeline, combined with SPADE normalisation, significantly improved the performance of the GAN models. To validate our findings, we conducted a clinical assessment in collaboration with radiologists. The synthesized CT images were evaluated for anatomical accuracy and diagnostic utility, demonstrating promising results in capturing critical pulmonary structures.

The second study addressed the limitations of current evaluation metrics in medical image synthesis, which often rely on global intensity measures that do not adequately reflect clinical relevance. We proposed a task-specific evaluation approach centered on segmentation metrics, such as the Dice coefficient and Normalized Surface Distance (NSD), to assess the reconstruction of fine anatomical structures like bronchi and intra-lung lesions. This shift towards task-specific metrics provided a more meaningful evaluation of the models' performance in a clinical context. Furthermore, we enhanced the synthesis process by integrating perceptual loss functions, which leverage feature representations from pre-trained networks to preserve high-level structural information. Additionally, to incorporate 3D contextual information, we adopted 2.5D approaches. By combining multi-axial fusion and stacked slice methods, we integrated spatial context without the computational burden of full 3D models. This hybrid approach improved the preservation of fine details and demonstrated superior performance in reconstructing complex lung anatomy.

In the third study, we recognized the need for a robust, scalable, and generalizable baseline for medical image synthesis, as existing state-of-the-art methods lacked standardization and broad applicability. We adapted the nnU-Net framework, originally designed for segmentation tasks, to suit the requirements of image synthesis. We demonstrated the effectiveness of our adapted nnU-Net on several public datasets and challenges, including pelvis MR to CT synthesis and brain local inpainting tasks. The model consistently matched or outperformed existing methods, showcasing its versatility and robustness. To encourage collaboration and further advancements in the field, we made our adapted

nnU-Net code publicly available. Finally, by understanding that precise reconstruction of anatomical structures is paramount in medical imaging, particularly in lung synthesis, we focused on enhancing the model's ability to capture fine details. We designed a novel loss function called the Anatomical Feature-Prioritized (AFP) loss, which emphasizes the accurate reconstruction of specific anatomical features by incorporating knowledge from pre-trained segmentation models. By introducing the AFP loss into our nnU-Net pipeline, we significantly improved the detailed MR to CT synthesis, achieving excellent results in thoracic imaging. We validated this approach through extensive task-based evaluations, including the segmentation and reconstruction of bronchi, intra-lung lesions, and extra-lung organs. The AFP loss effectively guided the model to prioritize critical structures, leading to synthetic images with high anatomical fidelity.

Beyond the immediate outcomes, this thesis prompts reflection on the broader implications of our work. Is the synthesis of CT images the ultimate goal, or does it serve as a stepping stone toward unlocking the full clinical potential of UTE MRI? We argue that while synthetic CT images offer significant value, they also pave the way for further advancements in non-invasive and patient-safe imaging practices. The synthetic images can be utilized to generate annotated databases based on existing CT segmentation models, thereby facilitating the development of segmentation, analysis, and quantification models directly applicable to UTE MRI data. Such models could assist radiologists in assessing disease severity and making more accurate diagnoses without the need for ionizing radiation.

In conclusion, this thesis makes novel contributions to the field of medical image synthesis, particularly in thoracic imaging, by demonstrating the feasibility and efficacy of MR to CT translation. Through the use of deep learning techniques, such as advanced loss functions and a focus on anatomical fidelity, we have addressed key challenges and improved the accuracy of pulmonary imaging. The collaboration between clinical and computational disciplines exemplifies the interdisciplinary efforts required to tackle complex healthcare challenges, with the potential to transform clinical practice and contribute to the advancement of precision medicine. The integration of deep learning with advanced evaluation methods holds great promise for improving patient outcomes and enhancing healthcare practices.

Perspectives

Following the contributions of this thesis, multiple directions for future research and development can be explored. These perspectives aim to enhance our methodologies, broaden the applications, and further refine evaluation metrics and techniques in medical imaging through deep learning.

Refinement of the Anatomical Feature-Prioritized (AFP) Loss Function

An immediate direction for improvement lies in refining the AFP loss function. While the AFP loss has been proven effective in guiding the model to prioritize critical anatomical structures during image synthesis, there is potential to enhance its adaptability and performance. Currently, the weights assigned to different feature maps within the AFP loss are static, which may not optimally reflect the varying importance of anatomical features

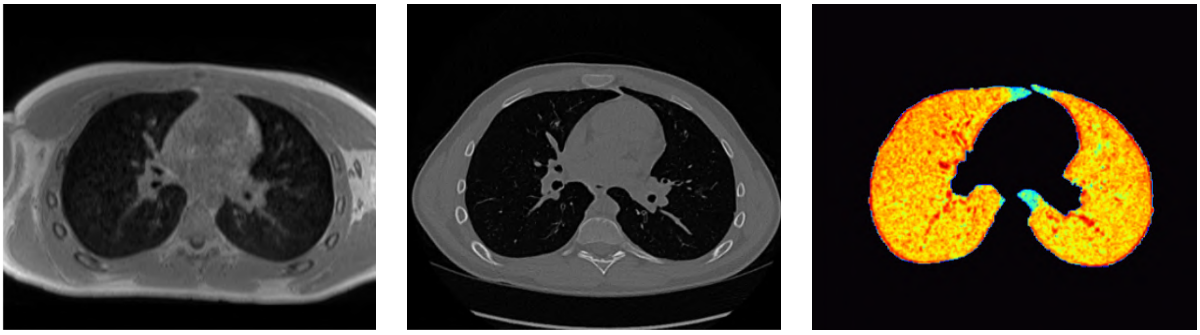
across different datasets or during different stages of training. We conducted preliminary experiments to fully understand the impact of different layers by assigning varied weights to individual feature maps within the AFP loss. These experiments provided insights into how specific layers contributed to the preservation of anatomical features, but the manual process highlighted the need for a more dynamic and automated approach.
Future work could involve implementing dynamic weighting strategies within the AFP loss, allowing the model to self-adjust these weights during training. This would enhance adaptability to different datasets and tasks, reducing manual tuning and optimizing performance across various anatomical regions and imaging modalities. Additionally, our work opens the opportunity to incorporate physical priors and acquisition parameters into the deep learning pipeline. By integrating information from the k-space domain of MRI data with the raw frequency and phase-encoded signals acquired during scans, we could provide the model with fundamental insights into the physics of image formation. Embedding acquisition-specific parameters or constraints derived from MRI physics may improve the fidelity of the synthesized images and reduce artifacts related to acquisition variability.
Validating the effectiveness of the AFP loss on different anatomical regions, such as brain or cardiac imaging, would further establish its generalizability in medical image synthesis. By applying our methods to diverse anatomical structures and integrating physical priors, we aim to develop more robust models adaptable to various clinical scenarios. This approach enhances technical performance and aligns with our broader goal of integrating domain-specific knowledge, both anatomical and physical, into deep learning frameworks to advance medical imaging.

Integration of 2.5D Approaches into the nnU-Net Framework

Another promising direction is the incorporation of 2.5D approaches into the nnU-Net pipeline. Our preliminary study on hybrid 2.5D models indicates that these approaches can effectively capture 3D contextual information while mitigating some of the challenges associated with 3D models. Building upon this, we are currently working on integrating 2.5D methods into the adapted nnU-Net pipeline. Early results from our experiments show that the 2.5D approaches generally outperform the 3D patch-based methods both quantitatively and qualitatively. The 3D patch-based methods often suffer from difficulties during image reconstruction, including edge effects, challenges in maintaining global consistency, and dependency on patch size. Moreover, full-volume 3D approaches are frequently not feasible due to memory constraints.
To validate these findings and enhance the robustness of our models, we propose to integrate 2.5D methods into the adapted nnU-Net pipeline. This integration would facilitate systematic comparisons between 2D, 2.5D, and 3D approaches within a consistent and reliable framework. Such studies would provide valuable insights into the most effective strategies for medical image synthesis, evaluating factors such as synthesis accuracy, computational efficiency, and the ability to preserve anatomical fidelity. The outcomes could guide the development of optimized models that balance performance and practicality, ultimately making advanced image synthesis techniques more accessible for clinical applications.

Incorporation of Error Maps into Synthesis Models

Accurate registration between MRI and CT images is a critical preprocessing step in our synthesis pipeline. However, the registration process is inherently prone to uncertainties and errors due to factors such as differences in acquisition parameters, patient positioning, and respiratory motion between modalities. Acknowledging that registration can yield multiple plausible solutions, we have developed an error map to quantify the inherent registration errors.
Our error map combines gradient field information and intensity differences within local patches, enabling it to capture poorly registered regions in a 3D spatial context. A key feature of this error map is its robustness to cross-contrast intensities, which is crucial in multimodal imaging where MRI and CT intensities differ significantly. By introducing a parameter that adjusts the relative importance of gradient field information, we prioritize spatial discrepancies over intensity differences. This approach allows the error map to effectively capture misalignments between registered images, making it well-suited for scenarios where traditional intensity-based comparisons are inadequate due to inherent contrast differences between modalities. Figure 4.11 illustrates an example of the error map following MR and CT registration, highlighting areas where differences are most pronounced.



Fixed MR | Registered CT | Registration error map

Figure 4.11: The error map highlights discrepancies between the fixed MR image and the registered CT image. Aligned regions appear in red, while areas with significant differences, such as the lung extremities, are in blue. These discrepancies, caused by differing respiratory phases during acquisition, are effectively captured by the error map.

The central idea is to incorporate this error map into the synthesis model so that it can account for registration errors during training. By adjusting the training gradients in proportion to the fidelity of the registration, the model can focus on learning from regions where the registration is reliable and be less influenced by areas with significant misalignment. We have already conducted preliminary experiments integrating the error map into our synthesis model, and initial results indicate that this approach enhances the robustness and accuracy of the model, particularly in the presence of imperfect registration.
Our error maps are specifically designed for 3D data, offering the potential to extend these benefits to full volumetric synthesis models. Integrating the error map into the nnU-Net pipeline or other advanced architectures shows promising improvements in the model's ability to handle registration uncertainties.

Advancing Evaluation Metrics in Medical Image Synthesis

A critical long-term perspective centers on the evaluation metrics used in medical imaging, particularly in image synthesis. The current state-of-the-art often relies on global intensity metrics such as Mean Absolute Error (MAE) and Structural Similarity Index Measure (SSIM), which may not adequately reflect the clinical utility of the synthesized images. These metrics can fail to capture the anatomical and functional accuracy required for diagnostic purposes.

There is an urgent need to develop and adopt evaluation metrics that align more closely with clinical requirements. Recent comprehensive studies, led by the Division of Intelligent Medical Systems at the German Cancer Research Center (DKFZ) have highlighted that flaws in machine learning algorithm validation are an underestimated global issue [154, 105]. In biomedical image analysis, current performance metrics often fail to reflect the domain's needs, hindering the translation of machine learning techniques into practice. These studies also propose several metrics and validation recommendations aimed at improving the reliability and clinical relevance of machine learning models in medical imaging. These studies represent a preliminary step, primarily addressing classification and segmentation tasks, leaving room for further development in areas like synthesis and registration.

Building on this, we have joined the preliminary study study initiated by the DKFZ team to discuss the use of metrics in medical image synthesis. Our collaboration aims to contribute to the development of standardized, clinically relevant evaluation frameworks that can be widely adopted in the medical imaging community. By focusing on metrics that assess anatomical accuracy, structural fidelity, and clinical relevance, we can improve the validation of synthesis models and facilitate their integration into clinical workflows.

Future work will involve participating in the development of guidelines and best practices for evaluating medical image synthesis algorithms. This includes advocating for the use of task-specific metrics, promoting transparency in reporting results, and encouraging the validation of models on diverse datasets. By contributing to this effort, we hope to enhance the reliability and credibility of machine learning applications in medical imaging.

Bibliography

[1] J. Iglesias, E. Konukoglu, D. Zikic, *et al.*, "Is synthesizing mri contrast useful for inter-modality analysis?," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, vol. 8149 of *Lecture Notes in Computer Science*, (Berlin, Heidelberg), Springer, Berlin, Heidelberg, 2013.

[2] S. A. Bobman, S. J. Riederer, J. N. Lee, *et al.*, "Cerebral magnetic resonance image synthesis," *American Journal of Neuroradiology*, vol. 6, no. 2, pp. 265–269, 1985.

[3] B. Fischl, D. H. Salat, A. J. van der Kouwe, *et al.*, "Sequence-independent segmentation of magnetic resonance images," *NeuroImage*, vol. 23, no. Suppl 1, pp. S69–S84, 2004.

[4] I. C. London, "Ixi dataset," 2014. This data is made available under the Creative Commons CC BY-SA 3.0 license.

[5] J. B. Warntjes, O. D. Leinhard, J. West, *et al.*, "Rapid magnetic resonance quantification on the brain: Optimization for clinical usage," *Magnetic Resonance in Medicine*, vol. 60, no. 2, pp. 320–329, 2008.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[7] A. Guimond, A. Roche, N. Ayache, *et al.*, "Three-dimensional multimodal brain warping using the demons algorithm and adaptive intensity corrections," *IEEE Transactions on Medical Imaging*, vol. 20, no. 1, pp. 58–69, 2001.

[8] D. J. Kroon and C. H. Slump, "Mri modalitiy transformation in demon registration," in *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 963–966, IEEE, June 2009.

[9] S. Roy, A. Carass, J. L. Prince, *et al.*, "Magnetic resonance image example-based contrast synthesis," *IEEE Transactions on Medical Imaging*, vol. 32, no. 12, pp. 2348–2363, 2013.

[10] N. Burgos *et al.*, "Attenuation correction synthesis for hybrid pet-mr scanners: Application to brain studies," *IEEE Transactions on Medical Imaging*, vol. 33, no. 12, pp. 2332–2341, 2014.

[11] N. Burgos *et al.*, "Robust ct synthesis for radiotherapy planning: Application to the head and neck region," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9350 of *Lecture Notes in Computer Science*, Springer, Cham, 2015.

[12] D. H. Ye, D. Zikic, B. Glocker, *et al.*, "Modality propagation: Coherent synthesis of subject-specific scans with data-driven regularization," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, pp. 606–613, Springer, Berlin, Heidelberg, 2013.

[13] F. Rousseau, P. A. Habas, and C. Studholme, "A supervised patch-based approach for human brain labeling," *IEEE Transactions on Medical Imaging*, vol. 30, no. 10, pp. 1852–1862, 2011.

[14] A. Jog, A. Carass, S. Roy, *et al.*, "Random forest regression for magnetic resonance image synthesis," *Medical Image Analysis*, vol. 35, pp. 475–488, 2017.

[15] T. Huynh, Y. Gao, J. Kang, *et al.*, "Estimating ct image from mri data using structured random forest and auto-context model," *IEEE Transactions on Medical Imaging*, vol. 35, no. 1, pp. 174–183, 2016.

[16] R. Vemulapalli, H. Van Nguyen, and S. Kevin Zhou, "Unsupervised cross-modal synthesis of subject-specific scans," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 630–638, IEEE, 2015.

[17] Y. LeCun, Y. Bengio, G. Hinton, *et al.*, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[18] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[19] A. Krizhevsky, I. Sutskever, G. E. Hinton, *et al.*, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, pp. 1097–1105, 2012.

[20] R. Li, W. Zhang, H. I. Suk, *et al.*, "Deep learning based imaging data completion for improved brain disease diagnosis," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, vol. 8652 of *Lecture Notes in Computer Science*, pp. 305–312, Springer, Cham, 2014.

[21] O. Ronneberger, P. Fischer, T. Brox, *et al.*, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351 of *Lecture Notes in Computer Science*, Springer, Cham, 2015.

[22] C. Zhao, A. Carass, J. Lee, *et al.*, "Whole brain segmentation and labeling from ct using synthetic mr images," in *Machine Learning in Medical Imaging – MLMI 2017*, vol. 10541 of *Lecture Notes in Computer Science*, Springer, Cham, 2017.

[23] X. Han, "Mr-based synthetic ct generation using a deep convolutional neural network method," *Medical Physics*, vol. 44, no. 4, pp. 1408–1419, 2017.

[24] P. Isola, J.-Y. Zhu, T. Zhou, *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1125–1134, IEEE, 2017.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680, Curran Associates, Inc., 2014.

[26] D. Pathak, P. Krähäenbühl, J. Donahue, *et al.*, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2536–2544, IEEE, 2016.

[27] D. Nie *et al.*, "Medical image synthesis with context-aware generative adversarial networks," in *Medical Image Computing and Computer Assisted Intervention - MICCAI 2017*, vol. 10435 of *Lecture Notes in Computer Science*, Springer, Cham, 2017.

[28] C. Han, H. Hayashi, L. Rundo, *et al.*, "Gan-based synthetic brain mr image generation," in *IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 734–738, IEEE, 2018.

[29] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[30] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, *et al.*, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, pp. 2256–2265, PMLR, June 2015.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[32] R. Rombach, A. Blattmann, D. Lorenz, *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, IEEE, 2022.

[33] W. H. Pinaya, P. D. Tudosiu, J. Dafflon, *et al.*, "Brain imaging generation with latent diffusion models," in *MICCAI Workshop on Deep Generative Models*, pp. 117–126, Springer Nature Switzerland, September 2022.

[34] G. Müller-Franzes, J. Niehues, F. Khader, *et al.*, "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, p. 12098, 2023.

[35] J. Wolleb, F. Bieder, R. Sandkühler, *et al.*, "Diffusion models for medical anomaly detection," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45, Springer Nature Switzerland, September 2022.

[36] M. Mirza *et al.*, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[37] X. Yi, E. Walia, P. Babyn, *et al.*, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[38] D. Nie, X. Cao, Y. Gao, *et al.*, "Estimating ct image from mri data using 3d fully convolutional networks," in *Deep learning and data labeling for medical applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, held in conjunction with MICCAI 2016*, pp. 170–178, Springer, Cham, 2016.

[39] P. Isola, J. Y. Zhu, T. Zhou, *et al.*, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, IEEE, 2017.

[40] Y. Wang, B. Yu, L. Wang, *et al.*, "3d conditional generative adversarial networks for high-quality pet image estimation at low dose," *Neuroimage*, vol. 174, pp. 550–562, 2018.

[41] A. Thummerer, E. van der Bijl, A. Galapon Jr, *et al.*, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical Physics*, vol. 50, no. 7, pp. 4664–4674, 2023.

[42] E. Huijben, M. Terpstra, A. Galapon Jr, *et al.*, "Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report," *arXiv preprint arXiv:2403.08447*, 2024.

[43] T. C. Wang, M. Y. Liu, J. Y. Zhu, *et al.*, "High-resolution image synthesis and semantic manipulation with conditional gans," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

[44] T. Park, M.-Y. Liu, T.-C. Wang, *et al.*, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, IEEE, 2019.

[45] L. Kong, C. Lian, D. Huang, *et al.*, "Breaking the dilemma of medical image-to-image translation," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 1964–1978, NeurIPS, 2021.

[46] B. Menze, A. Jakab, S. Kalpathy-Craler, *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[47] J. Y. Zhu, T. Park, P. Isola, *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, IEEE, 2017.

[48] J. Wolterink, A. Dinkla, M. Savenije, *et al.*, "Deep mr to ct synthesis using unpaired data," in *Simulation and Synthesis in Medical Imaging. SASHIMI 2017*, vol. 10557 of *Lecture Notes in Computer Science*, p. 2, Springer, Cham, 2017.

[49] Y. Hiasa, Y. Otake, M. Takao, *et al.*, "Cross-modality image synthesis from unpaired data using cyclegan: Effects of gradient consistency loss and training data size," in *Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018*, pp. 31–41, Springer International Publishing, 2018.

[50] R. Chen, W. Huang, B. Huang, *et al.*, "Reusing discriminators for encoding: Towards unsupervised image-to-image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8168–8177, IEEE, 2020.

[51] M.-Y. Liu, T. Breuel, J. Kautz, *et al.*, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems*, vol. 30, Curran Associates, Inc., 2017.

[52] T. Park, A. A. Efros, R. Zhang, *et al.*, "Contrastive learning for unpaired image-to-image translation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345, Springer International Publishing, 2020.

[53] J. P. Cohen, M. Luck, S. Honari, *et al.*, "Distribution matching losses can hallucinate features in medical image translation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pp. 529–536, Springer International Publishing, 2018.

[54] Z. Wang, A. C. Bovik, H. R. Sheikh, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, April 2004.

[55] A. Mason, J. Rioux, S. E. Clarke, *et al.*, "Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of mr images," *IEEE Transactions on Medical Imaging*, vol. 39, pp. 1064–1072, April 2020.

[56] A. Gretton, K. M. Borgwardt, M. J. Rasch, *et al.*, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[57] M. Heusel, H. Ramsauer, T. Unterthiner, *et al.*, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proceedings of the 31st International Conference on NIPS*, pp. 6629–6640, 2017.

[58] L. Xiang, Q. Wang, D. Nie, *et al.*, "Deep embedding convolutional neural network for synthesizing ct image from t1-weighted mr image," *Medical Image Analysis*, vol. 47, pp. 31–44, 2018.

[59] M. Dohmen, T. Truong, I. M. Baltruschat, *et al.*, "Five pitfalls when assessing synthetic medical images with reference metrics," in *Deep Generative Models. DGM4MICCAI 2024*, vol. 15224, p. 15, Springer, Cham, 2025.

[60] M. Bińkowski, D. J. Sutherland, M. Arbel, *et al.*, "Demystifying mmd gans," *arXiv preprint arXiv:1801.01401*, 2018.

[61] M. J. Chong and D. Forsyth, "Effectively unbiased fid and inception score and where to find them," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, IEEE/CVF, 2020.

[62] K. Armanious, C. Jiang, M. Fischer, *et al.*, "Medgan: Medical image translation using gans," *Computerized Medical Imaging and Graphics*, vol. 79, p. 101684, 2020.

[63] Y. Lei, J. Harms, T. Wang, *et al.*, "Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks," *Medical Physics*, vol. 46, no. 8, pp. 3565–3581, 2019.

[64] C. Wang, G. Macnaught, G. Papanastasiou, *et al.*, "Unsupervised learning for cross-domain medical image synthesis using deformation invariant cycle consistency networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2018*, Springer, Cham, 2018.

[65] K. M. Johnson, S. B. Fain, M. L. Schiebler, *et al.*, "Optimized 3d ultrashort echo time pulmonary mri," *Magnetic Resonance in Medicine*, vol. 70, no. 5, pp. 1241–1250, 2013.

[66] K. Weiss, T. M. Khoshgoftaar, D. Wang, *et al.*, "A survey of transfer learning," *Journal of Big Data*, vol. 3, no. 1, p. 9, 2016.

[67] D. M. Hansell, A. A. Bankier, H. MacMahon, *et al.*, "Fleischner society: glossary of terms for thoracic imaging," *Radiology*, vol. 246, no. 3, pp. 697–722, 2008.

[68] G. Dournes, D. Grodzki, B. Mädler, *et al.*, "Quiet submillisecond pulmonary mr imaging with petra: Comparison with hrct in 50 subjects," *Radiology*, vol. 276, no. 1, pp. 258–265, 2015.

[69] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, IEEE, 2017.

[70] G. Dournes, J. Yazbek, W. Benhassen, *et al.*, "3d ultrashort echo-time mri of the lung using stack-of-spirals and spherical k-space coverages: Evaluation in healthy volunteers and parenchymal diseases: Lung mri with 3d ute spiral vibe sequence," *Journal of Magnetic Resonance Imaging*, vol. 48, 2018.

[71] J. C. Reinhold, B. E. Dewey, A. Carass, *et al.*, "Evaluating the impact of intensity normalization on mr image synthesis," in *Proceedings of SPIE–the International Society for Optical Engineering*, vol. 10949, SPIE, March 2019.

[72] D. P. Shamonin, E. E. Bron, B. P. Lelieveldt, *et al.*, "Fast parallel image registration on cpu and gpu for diagnostic classification of alzheimer's disease," *Front Neuroinformatics*, vol. 7, p. 50, January 16 2014.

[73] P. Thevenaz, T. Blu, and M. Unser, "Interpolation revisited [medical images application]," *IEEE Transactions on Medical Imaging*, vol. 19, no. 7, pp. 739–758, 2000.

[74] B. C. Lowekamp, D. T. Chen, L. Ibáñez, *et al.*, "The design of simpleitk," *Front Neuroinformatics*, vol. 7, p. 45, December 30 2013.

[75] M. P. Heinrich, M. Jenkinson, M. Bhushan, *et al.*, "Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration," *Medical Image Analysis*, vol. 16, no. 7, pp. 1423–1435, 2012.

[76] B. Denis de Senneville, C. Zachiu, M. Ries, and C. Moonen, "Evolution: an edge-based variational method for non-rigid multi-modal image registration," *Physics in Medicine and Biology*, vol. 61, no. 20, p. 7377, 2016.

[77] B. B. Avants, C. L. Epstein, M. Grossman, *et al.*, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[78] S. Klein, M. Staring, K. Murphy, *et al.*, "elastix: a toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, no. 1, pp. 196–205, 2010.

[79] M. Modat, G. R. Ridgway, Z. A. Taylor, *et al.*, "Fast free-form deformation using graphics processing units," *Computer Methods and Programs in Biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[80] M. P. Heinrich, M. Jenkinson, B. W. Papież, *et al.*, "Towards real-time multimodal fusion for image-guided interventions using self-similarities," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 187–194, Springer, Berlin, Heidelberg, 2013.

[81] G. Balakrishnan, A. Zhao, M. R. Sabuncu, *et al.*, "Voxelmorph: a learning framework for deformable medical image registration," *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1788–1800, 2019.

[82] Y. Hu, M. Modat, E. Gibson, *et al.*, "Deepreg: a deep learning toolkit for medical image registration," *Journal of Open Source Software*, vol. 5, no. 55, p. 2705, 2020.

[83] H. Siebert, L. Hansen, and M. Heinrich, "Fast 3d registration with accurate optimisation and little learning for learn2reg 2021," in *Biomedical Image Registration, Domain Generalisation and Out-of-Distribution Analysis - MICCAI 2021 Challenges: MIDOG 2021, MOOD 2021, and Learn2Reg 2021*, vol. 13166, pp. 174–179, Springer, 2021.

[84] H. Siebert, C. Großbröhmer, L. Hansen, *et al.*, "Convexadam: Self-configuring dual-optimisation-based 3d multitask medical image registration," *IEEE Transactions on Medical Imaging*, 2024.

[85] A. Hering, L. Hansen, T. C. Mok, *et al.*, "Learn2reg: comprehensive multi-task medical image registration challenge, dataset and evaluation in the era of deep learning," *IEEE Transactions on Medical Imaging*, vol. 42, no. 3, pp. 697–712, 2022.

[86] Y. Fu, Y. Lei, T. Wang, *et al.*, "Deep learning in medical image registration: a review," *Physics in Medicine & Biology*, vol. 65, no. 20, p. 20TR01, 2020.

[87] M. Irani and P. Anandan, "Robust multi-sensor image alignment," *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, pp. 959–966, 1998.

[88] L. Lafitte, C. Zachiu, L. G. W. Kerkmeijer, *et al.*, "Accelerating multi-modal image registration using a supervoxel-based variational framework," *Physics in Medicine and Biology*, vol. 63, no. 23, p. 235009, 2018.

[89] G. E. Christensen, H. J. Johnson, *et al.*, "Consistent image registration," *IEEE Transactions on Medical Imaging*, vol. 20, no. 7, pp. 568–582, 2001.

[90] J. Hofmanninger, F. Prayer, J. Pan, *et al.*, "Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem," *European Radiology Experimental*, 2020.

[91] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, ICLR, 2015.

[92] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *4th International Conference on Learning Representations (ICLR)*, ICLR, 2016.

[93] F. Isensee, P. F. Jaeger, S. A. Kohl, *et al.*, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021.

[94] J. Modersitzki, *Numerical Methods for Image Registration*. Oxford University Press, 2004.

[95] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4ITK: improved N3 bias correction," *IEEE Trans Med Imaging*, vol. 29, pp. 1310–1320, Jun 2010.

[96] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," *IEEE Trans Med Imaging*, vol. 19, pp. 143–150, Feb 2000.

[97] E. Seeram, *Computed Tomography: Physical Principles, Clinical Applications, and Quality Control*. Elsevier Health Sciences, 4th ed., 2015.

[98] H. Zhao, O. Gallo, I. Frosio, *et al.*, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2017.

[99] S. Lin, C. Zhang, and Y. Yang, "A pluggable single-image super-resolution algorithm based on second-order gradient loss," *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, vol. 3, no. 4, p. 100148, 2023.

[100] A. Tsitsulin, M. Munkhoeva, D. Mottin, *et al.*, "The shape of data: Intrinsic distance for data distributions," in *ICLR'2020*, 2020.

[101] A. Borji, "Pros and cons of gan evaluation measures," *Computer Vision and Image Understanding*, vol. 179, pp. 41–65, 2019.

[102] A. Wang, T. Tam, H. Poon, *et al.*, "Naviairway: a bronchiole-sensitive deep learning-based airway segmentation pipeline for planning of navigation bronchoscopy," *Transactions on Medical Imaging*, 2022.

[103] M. Zhang, Y. Wu, H. Zhang, *et al.*, "Multi-site, multi-domain airway tree modeling," *Medical Image Analysis*, vol. 90, p. 102957, 2023.

[104] S. Nikolov, S. Blackwell, A. Zverovitch, *et al.*, "Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study," *Journal of Medical Internet Research*, vol. 23, no. 7, p. e26151, 2021.

[105] L. Maier-Hein, A. Reinke, P. Godau, and et al., "Metrics reloaded: recommendations for image analysis validation," *Nature Methods*, vol. 21, pp. 195–212, 2024.

[106] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, pp. 694–711, Springer, Cham, 2016.

[107] X. Chen, X. Tian, J. He, *et al.*, "Perceptual adversarial networks for image-to-image transformation," *arXiv preprint arXiv:1706.09138*, 2017.

[108] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[109] A. Longuefosse, G. Dournes, I. Benlala, *et al.*, "Lung ct synthesis using gans with conditional normalization on registered ultrashort echo-time mri," in *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, 2023.

[110] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," *arXiv preprint arXiv:1603.07285*, 2016.

[111] G. Huang, Z. Liu, L. Van Der Maaten, *et al.*, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, IEEE, 2017.

[112] T. Karras, S. Laine, T. Aila, *et al.*, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, IEEE, 2019.

[113] D. Shen, G. Wu, and H. I. Suk, "Deep learning in medical image analysis," *Annual Review of Biomedical Engineering*, vol. 19, pp. 221–248, 2017.

[114] L. A. Gatys, A. S. Ecker, M. Bethge, *et al.*, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423, IEEE, 2016.

[115] K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE, 2016.

[116] Z. Liu, H. Mao, C.-Y. Wu, *et al.*, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, IEEE, 2022.

[117] X. Mei, Z. Liu, P. M. Robson, *et al.*, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, p. e210259, 2022.

[118] J. Zhang, Y. Xie, Q. Wu, *et al.*, "Medical image classification using synergic deep learning," *Medical Image Analysis*, vol. 54, pp. 10–19, 2019.

[119] D. Koguciuk, E. Arani, B. Zonooz, *et al.*, "Perceptual loss for robust unsupervised homography estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4274–4293, IEEE/CVF, 2021.

[120] A. Andonian, T. Park, B. Russell, *et al.*, "Contrastive feature loss for image prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1934–1943, 2021.

[121] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," *arXiv preprint arXiv:2002.05709*, 2020.

[122] D. Juodelyte, A. Jiménez-Sánchez, and V. Cheplygina, "Revisiting hidden representations in transfer learning for medical imaging," *arXiv preprint arXiv:2302.08272*, 2023.

[123] S. K. Zhou, H. Greenspan, C. Davatzikos, *et al.*, "A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises," *Proceedings of the IEEE*, vol. 109, no. 5, pp. 820–838, 2021.

[124] Y. Zhang, Q. Liao, L. Ding, *et al.*, "Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5 d solutions," *Computerized Medical Imaging and Graphics*, vol. 99, p. 102088, 2022.

[125] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, IEEE, 2016.

[126] S. P. Singh, L. Wang, S. Gupta, *et al.*, "3d deep learning on medical images: A review," *Sensors*, vol. 20, p. 5097, 2020.

[127] H. R. Roth, L. Lu, A. Seff, *et al.*, "A new 2.5d representation for lymph node detection using random sets of deep convolutional neural network observations," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, (Cham), pp. 520–527, Springer International Publishing, 2014.

[128] Y. Xia, L. Xie, F. Liu, *et al.*, "Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, vol. 11073, p. 51, Springer, Cham, 2018.

[129] Q. Yu, Y. Xia, L. Xie, *et al.*, "Thickened 2d networks for efficient 3d medical image segmentation," *arXiv preprint arXiv:1904.01150*, 2019.

[130] A. Prasoon, K. Petersen, C. Igel, *et al.*, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*, vol. 16, pp. 246–253, Springer, Cham, 2013.

[131] H. Zheng, L. Qian, Y. Qin, *et al.*, "Improving the slice interaction of 2.5d cnn for automatic pancreas segmentation," *Medical Physics*, vol. 47, pp. 5543–5554, November 2020.

[132] H. Zhang, A. M. Valcarcel, R. Bakshi, *et al.*, "Multiple sclerosis lesion segmentation with tiramisu and 2.5d stacked slices," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*, vol. 11766 of *Lecture Notes in Computer Science*, pp. 338–346, Springer, Cham, 2019.

[133] F. Isensee, P. F. Jäger, P. M. Full, *et al.*, "nnu-net for brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part II 6*, pp. 118–132, Springer International Publishing, 2021.

[134] F. Isensee, T. Wald, C. Ulrich, *et al.*, "nnu-net revisited: A call for rigorous validation in 3d medical image segmentation," *arXiv preprint arXiv:2404.09556*, 2024.

[135] A. Longuefosse, J. Raoul, and I. e. a. Benlala, "Generating high-resolution synthetic ct from lung mri with ultrashort echo times: initial evaluation in cystic fibrosis," *Radiology*, vol. 308, no. 1, p. e230052, 2023.

[136] F. Kofler, F. Meissen, F. Steinbauer, *et al.*, "The brain tumor segmentation (brats) challenge 2023: Local synthesis of healthy brain tissue via inpainting," *arXiv preprint arXiv:2305.08992*, 2023.

[137] A. Lugmayr, M. Danelljan, A. Romero, *et al.*, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, IEEE/CVF, 2022.

[138] A. Durrer, J. Wolleb, F. Bieder, *et al.*, "Denoising diffusion models for 3d healthy brain tissue inpainting," *arXiv preprint arXiv:2403.14499*, 2024.

[139] A. P. Aitken, C. Ledig, L. Theis, *et al.*, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv preprint arXiv:1707.02937*, 2017.

[140] A. Odena, V. Dumoulin, and C. Olah, "Deconvolution and checkerboard artifacts," *Distill*, vol. 1, no. 10, p. e3, 2016.

[141] T. M. Consortium, "Project monai," Dec 2020.

[142] R. Zhu, X. Zhang, H. Pang, *et al.*, "Advancing brain tumor inpainting with generative models," *arXiv preprint arXiv:2402.01509*, 2024.

[143] W. Pinaya, M. Graham, and E. e. a. Kerfoot, "Generative ai for medical imaging: Extending the monai framework," *arXiv:2307.15208*, 2023.

[144] H. Shin, N. Tenenholtz, and J. e. a. Rogers, "Medical image synthesis for data augmentation and anonymization using generative adversarial networks," in *Simulation and Synthesis in Medical Imaging* (A. Gooya, O. Goksel, I. Oguz, and N. Burgos, eds.), vol. 11037 of *Lecture Notes in Computer Science*, Springer, 2018.

[145] V. Fernandez and et al., "Can segmentation models be trained with fully synthetically generated data?," in *Simulation and Synthesis in Medical Imaging* (C. Zhao, D. Svoboda, and J. e. a. Wolterink, eds.), vol. 13570 of *Lecture Notes in Computer Science*, Springer, 2022.

[146] H. Yang, J. Sun, and A. e. a. Carass, "Unsupervised mr-to-ct synthesis using structure-constrained cyclegan," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 4249–4261, 2020.

[147] S. Kaushik, M. Bylund, and C. e. a. Cozzini, "Region of interest focused mri to synthetic ct translation using regression and classification multi-task network," *arXiv preprint arXiv:2203.16288*, 2022.

[148] H. Emami, M. Dong, and S. e. a. Nejad-Davarani, "Sa-gan: Structure-aware gan for organ-preserving synthetic ct generation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, vol. 12906 of *Lecture Notes in Computer Science*, Springer, Cham, 2021.

[149] A. Longuefosse, E. Le Bot, and B. e. a. Denis de Senneville, "Adapted nnu-net: A robust baseline for cross-modality synthesis and medical image inpainting," in *International Workshop on Simulation and Synthesis in Medical Imaging*, October 2024. Accepted.

[150] A. I. H. Bouzid, B. D. Denis de Senneville, and F. e. a. Baldacci, "Ct evaluation of 2d and 3d holistic deep learning methods for the volumetric segmentation of airway lesions," *arXiv preprint arXiv:2403.08042*, 2024.

[151] J. Wasserthal, H.-C. Breit, and M. e. a. Meyer, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, 2023.

[152] A. Longuefosse, B. D. De Senneville, G. Dournes, *et al.*, "On the use of perceptual loss for fine structure generation: Illustration on lung mr to ct synthesis," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5, IEEE, 2024.

[153] N. Tajbakhsh, L. Jeyaseelan, Q. Li, and et al., "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[154] A. Reinke, M. D. Tizabi, M. Baumgartner, *et al.*, "Understanding metric-related pitfalls in image analysis validation," *Nature Methods*, vol. 21, pp. 182–194, 2024.