| | |
|---|---|
| **Titre:** Title: | Combined Image Synthesis and Deformable Registration Framework for MRI-Based Adaptive Radiotherapy |
| **Auteur:** Author: | Shima Sargordi |
| **Date:** | 2025 |
| **Type:** | Mémoire ou thèse / Dissertation or Thesis |
| **Référence:** Citation: | Sargordi, S. (2025). Combined Image Synthesis and Deformable Registration Framework for MRI-Based Adaptive Radiotherapy [Master's thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/71669/ |

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

# Combined Image Synthesis and Deformable Registration Framework for MRI-Based Adaptive Radiotherapy

**SHIMA SARGORDI**
Département de génie informatique et génie logiciel

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie informatique

Novembre 2025

**POLYTECHNIQUE MONTRÉAL**
affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Combined Image Synthesis and Deformable Registration Framework for MRI-Based Adaptive Radiotherapy**

présentée par **Shima SARGORDI**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment acceptée par le jury d'examen constitué de :

**Herve LOMBAERT**, président
**Samuel KADOURY**, membre et directeur de recherche
**Omar Abdel WAHAB**, membre

# DEDICATION

*To all my beloved family and friends*

## ACKNOWLEDGEMENTS

I am grateful to my supervisor, Professor Samuel Kadoury, for his guidance and support throughout this research. His expertise and mentorship have been invaluable.

My gratitude extends to Zeinab Abboud, a PhD fellow in our group, from whom I learned greatly. Her guidance and generosity in sharing knowledge and support were invaluable.

I would like to thank my fellows at the MediCAL laboratory for their support and motivation during my master's studies. I am thankful to William Le, Gautier Henique, David Grajales Lopera, Emilie Ouraou, Amine El Foraici, Ralph Saber, and Aloys Portafaix.

Finally, I am grateful to my parents, my brother Maziar, and my sister Beheshteh for their unwavering love and support.

# RÉSUMÉ

Le cancer de la tête et du cou (CTC) demeure l'un des cancers les plus complexes à traiter par radiothérapie, en raison de son anatomie hétérogène et de la proximité des tumeurs avec des organes à risque (OAR). La radiothérapie adaptative (ART) vise à compenser les variations anatomiques quotidiennes observées au cours du traitement en ajustant le plan selon l'anatomie du jour, afin d'améliorer la précision dosimétrique, de réduire la toxicité et de maintenir une couverture optimale de la tumeur. Dans ce contexte, l'imagerie par résonance magnétique (IRM) offre un contraste supérieur des tissus mous sans rayonnement ionisant, ce qui améliore le contourage et la planification. Le scanner (CT) reste essentiel pour le calcul de la dose, tandis que la tomographie volumique conique (CBCT), acquise avant chaque fraction, permet de suivre les changements anatomiques et d'adapter le plan de traitement. Cependant, la CBCT souffre d'artefacts, de bruit et d'unités Hounsfield non uniformes, ce qui limite son utilisation directe pour le calcul de dose. Un flux de travail ART basé uniquement sur l'IRM permettrait de réduire la dépendance au CT de planification, mais cela exige deux capacités essentielles : (i) une synthèse IRM–CT fiable pour générer des informations équivalentes au CT en vue du calcul de dose et (ii) un enregistrement déformable précis pour aligner le CT synthétique (sCT) sur la CBCT quotidienne.

Cette thèse présente une approche en deux étapes basée sur l'apprentissage profond pour la registration déformable entre l'IRM et la CBCT. La première étape concerne la génération d'images synthétiques, où plusieurs réseaux ont été explorés : un modèle de base de type Conditional Generative Adversarial Network (CGAN), une version améliorée intégrant une contrainte de cohérence de caractéristiques (FCGAN), ainsi qu'un modèle de diffusion transformeur MRI-to-CT (MC-IDDPM) offrant une meilleure restitution du contraste et des structures anatomiques. L'objectif est de produire des sCT présentant une fidélité anatomique et une cohérence d'intensité suffisantes pour être utilisées dans les étapes suivantes, notamment le calcul de dose, la segmentation et la registration déformable, éléments clés de la planification et du suivi du traitement. La seconde étape de la méthode se concentre sur la registration déformable du pCT ou du sCT vers les CBCT acquis quotidiennement, afin d'obtenir un alignement spatial précis, en particulier au niveau des régions tumorales et des organes à risque.

La synthèse a été évaluée à la fois sur l'ensemble de données cérébrales SynthRad2023 et sur une cohorte institutionnelle de 338 patients HNC avec IRM-TDM appariée et CBCT longitudinale, tandis que l'enregistrement a été évalué uniquement sur la cohorte HNC où les

trois modalités étaient disponibles. Sur les données HNC internes, le FCGAN a amélioré la fidélité structurelle par rapport au CGAN de base, réduisant le MAE de 29,0% et augmentant le PCC, le PSNR et le SSIM de 0,6%, 7,4% et 1,1%, respectivement. Par rapport au FCGAN, le MC-IDDPM a encore réduit le MAE de 22,7% et a produit des gains de 0,3% en PCC et de 8,1% en PSNR, indiquant une meilleure précision structurelle. Concernant le chevauchement structurel, le MC-IDDPM a obtenu un Dice supérieur de 1,2% sur le crâne, mais inférieur de 1,5% sur la moelle épinière et de 5,4% sur chaque parotide par rapport au FCGAN. Par rapport au CGAN, le FCGAN a augmenté le Dice de 6,5% (moelle épinière), de 1,2% (crâne) et de 5,7% (parotides gauche/droite). Globalement, le FCGAN a favorisé la fidélité des tissus mous (moelle épinière, parotides), tandis que le MC-IDDPM s'est avéré supérieur pour l'anatomie osseuse. Le MC-IDDPM a également réduit la différence de dose MAE et DVH de 1,3% supplémentaire par rapport au FCGAN, et a surpassé le FCGAN et le CGAN sur toutes les mesures de similarité d'image (MAE, PCC, PSNR, SSIM) sur SynthRad2023.

Lors de l'étape d'enregistrement déformable, le VM-XMorpher à transformateur a fourni la correspondance géométrique la plus cohérente entre la tomodensitométrie (CT) et la CBCT. Il a atteint la plus grande similarité structurelle (SSIM = $0{,}92 \pm 0{,}002$) tout en obtenant simultanément la plus faible erreur de repère (TRE = $0{,}67 \pm 0{,}37$ mm). Par rapport aux autres approches, le VM-XMorpher a réduit la TRE de 16,2% par rapport au VoxelMorph, de 41,2% par rapport au recalage B-spline, de 37,4% par rapport au PC-Reg-RT et de 22,1% par rapport au PC-XMorpher. Il a également amélioré la SSIM de 1,1% par rapport au VoxelMorph, de 3,4% par rapport au B-spline et de 8,2% par rapport à l'initialisation rigide/affine.

Dans notre pipeline final combinant la synthèse IRM-vers-CT et la registration déformable sCT-vers-CBCT, les sCT générés par le modèle MC-IDDPM ont été utilisés pour l'étape de registration, ce modèle ayant démontré les meilleures performances en termes de similarité d'image et d'évaluation dosimétrique. Sur l'ensemble de test HNC, MC-IDDPM a atteint une MAE de $0.040 \pm 0.018$, un PCC de $0.939 \pm 0.031$, un PSNR de $25.01 \pm 2.6$ et un SSIM de $0.92 \pm 0.024$, avec des indices de Dice de $0.67 \pm 0.086$ pour la moelle épinière, $0.49 \pm 0.16$ pour la parotide gauche, $0.53 \pm 0.12$ pour la parotide droite et $0.87 \pm 0.08$ pour le crâne. Le modèle a également obtenu une MAE dosimétrique de $1.53 \pm 0.42$ et une différence DVH de $0.758 \pm 0.50$ avec TTA, surpassant les réseaux CGAN et FCGAN. À partir de ces sCT, le modèle VM-XMorpher a atteint un NCC de $0.92 \pm 0.016$ et un SSIM de $0.92 \pm 0.019$ sur les paires sCT–CBCT déformées de l'ensemble de test.

Dans l'ensemble, ces travaux démontrent qu'une synthèse IRM-CT de haute qualité, combinée à un enregistrement déformable CT-CBCT robuste, peut soutenir un flux de travail ART basé sur l'IRM. Le pipeline proposé ouvre la voie à une réduction de la dépendance au

CT de planification tout en maintenant la précision nécessaire pour une planification adaptative du traitement. Cette étude met en évidence le potentiel des modèles de synthèse et d'enregistrement d'images basés sur l'apprentissage profond pour renforcer la personnalisation des traitements et améliorer la qualité de vie des patients atteints d'un cancer de la tête et du cou soumis à une radiothérapie.

# ABSTRACT

Head and neck cancer (HNC) is among the most challenging malignancies to treat with radiotherapy due to its heterogeneous anatomy and the close proximity of tumors to organs at risk (OARs). Adaptive radiotherapy (ART) has emerged to address day-to-day anatomical changes during treatment by adjusting the plan to the patient's current anatomy, with the goals of improving dose accuracy, reducing treatment-related toxicity, and maintaining effective tumor control. In this workflow, magnetic resonance imaging (MRI) provides superior soft-tissue contrast without ionizing radiation, enabling better diagnosis and treatment planning; computed tomography (CT) remains indispensable for dose calculation; and cone-beam CT (CBCT), acquired before each fraction, offers daily volumetric imaging to monitor anatomy and align the pCT and its dose plan to the day's anatomy. However, CBCT suffers from noise, artifacts, and inconsistent Hounsfield Units (HU), which limit direct dose calculation. A fully MRI-based ART pathway could reduce reliance on pCT, but this requires (i) reliable MRI-to-CT synthesis to supply CT-equivalent information for dose calculation and (ii) accurate deformable registration to align the synthetic CT (sCT) with the daily CBCT.

This thesis proposes a two-stage deep learning pipeline for MRI-to-CBCT deformable registration. In the first stage, multiple synthesis networks were investigated, including a baseline Conditional Generative Adversarial Network (CGAN), a Feature-Consistency CGAN (FCGAN) incorporating a feature consistency loss term, and a transformer-based MRI-to-CT Improved Denoising Diffusion Probabilistic Model (MC-IDDPM) for synthesizing CT from MRI. The goal is to generate sCTs with sufficient anatomical and intensity fidelity to be suitable for subsequent tasks such as dose calculation, segmentation, and deformable registration, which are essential in treatment planning and delivery. In the second deformable registration stage, the objective was to align the planning CT (pCT) or the sCT to the daily CBCTs, achieving fine spatial alignment, particularly within the tumor and OAR regions.

Synthesis was evaluated on both the SynthRad2023 brain dataset [1] and an institutional HNC cohort comprising 338 patients with paired MRI, CT, and serial CBCTs, while registration was assessed exclusively on the HNC dataset, where all three modalities were available. On HNC cohort, FCGAN enhanced structural fidelity over the baseline CGAN with an 29.0% Mean Absolute Error (MAE) reduction and respective increases in Pearson Correlation Coefficient (PCC), Peak Signal to Noise Ratio (PSNR), and Structural Similarity Index Measure (SSIM) of 0.6%, 7.4%, and 1.1%. MC-IDDPM reduced MAE by 22.7% and increased PCC and PSNR by 0.3% and 8.1%, respectively, compared with FCGAN, demon-

strating improved structural accuracy. Compared with FCGAN, MC-IDDPM yields a 1.2% higher Dice for the skull, but is lower by 1.5% for the spinal cord and by 5.4% for each parotid (left and right). Relative to CGAN, FCGAN raises the Dice by 6.5% for the spinal cord, 1.2% for the skull, and 5.7% for each parotid gland. In summary, FCGAN performs better on soft tissues (spinal cord and parotids), whereas MC-IDDPM is superior for bony anatomy. With test-time augmentation, MC-IDDPM achieved 1.3% reduction in dose MAE and dose–volume histogram (DVH) difference relative to FCGAN, reflecting enhanced dosimetric consistency. MC-IDDPM also outperformed FCGAN and CGAN across all image similarity metrics, including MAE, PCC, PSNR, and SSIM, on the SynthRad2023 brain dataset.

In the second deformable registration stage, VM-XMorpher achieved the highest SSIM ($0.92\pm 0.002$) and the lowest target registration error ($0.67\pm0.37$ mm) across all methods, indicating superior CT-to-CBCT alignment. VM-XMorpher achieved target registration error reductions of 16.2% versus VoxelMorph, 41.2% versus B-spline, 37.4% versus PC-Reg-RT, and 22.1% versus PC-XMorpher. It also yielded SSIM gains of 1.1%, 3.4%, and 8.2% over VoxelMorph, B-spline, and affine/rigid initialization, respectively.

In our final pipeline of MRI-to-CT synthesis and sCT-to-CBCT registration, the sCTs from MC–IDDPM were used for the deformable registration stage, as this model achieved the best image similarity and dose evaluation metrics. On the HNC test set, MC-IDDPM reached an MAE of $0.040\pm0.018$, PCC of $0.939\pm0.031$, PSNR of $25.01\pm2.6$, and SSIM of $0.92\pm0.024$, with Dice scores of $0.67 \pm 0.086$ (spinal cord), $0.49 \pm 0.16$ (left parotid), $0.53 \pm 0.12$ (right parotid), and $0.87 \pm 0.08$ (skull). It also achieved a dose MAE of $1.53 \pm 0.42$ and a DVH difference of $0.758 \pm 0.50$ with TTA, outperforming both CGAN and FCGAN. Using these sCTs, VM-XMorpher attained an NCC of $0.92 \pm 0.016$ and SSIM of $0.92 \pm 0.019$ on the deformed sCT–CBCT pairs from the test set.

Together, these findings demonstrate that MRI-to-CT synthesis combined with deformable CT-to-CBCT registration can support a promising MRI-based ART workflow. The proposed pipeline establishes a pathway toward reducing reliance on pCT while maintaining the accuracy required for adaptive treatment planning. This work highlights the potential of deep generative and registration models to improve treatment personalization and patient quality of life in HNC radiotherapy.

# TABLE OF CONTENTS

DEDICATION . . . . . iii

ACKNOWLEDGEMENTS . . . . . iv

RÉSUMÉ . . . . . v

ABSTRACT . . . . . viii

LIST OF TABLES . . . . . xiii

LIST OF FIGURES . . . . . xiv

LIST OF SYMBOLS AND ACRONYMS . . . . . xviii

CHAPTER 1 INTRODUCTION . . . . . 1
1.1 Outline of the thesis . . . . . 2

CHAPTER 2 BACKGROUND . . . . . 4
2.1 Head and Neck Cancer . . . . . 4
2.1.1 Anatomical Structures in HNC . . . . . 4
2.1.2 Adaptive Radiotherapy in HNC . . . . . 5
2.2 Medical Imaging for HNC . . . . . 5
2.2.1 Magnetic Resonance Imaging (T1-weighted) . . . . . 6
2.2.2 Computed Tomography . . . . . 7
2.2.3 Cone Beam Computed Tomography . . . . . 8
2.2.4 Dose Planning . . . . . 9
2.3 The workflow for ART . . . . . 10

CHAPTER 3 LITERATURE REVIEW . . . . . 12
3.1 Deep Learning Concepts . . . . . 12
3.1.1 Multilayer Neural Network . . . . . 12
3.1.2 Convolutional Neural Networks . . . . . 13
3.1.3 U-Net . . . . . 14
3.2 Medical Image Synthesis . . . . . 15
3.2.1 Classical Methods . . . . . 16
3.2.2 Deep Learning Methods . . . . . 16
3.2.3 Image Similarity Metrics for Synthesis Evaluation . . . . . 23

3.2.4 Dose evaluation metrics . . . 25
3.3 Medical Image Registration . . . 26
3.3.1 Rigid Registration . . . 26
3.3.2 Affine Registration . . . 26
3.3.3 Deformable (Non-Rigid) Registration . . . 26
3.3.4 Registration Evaluation metrics . . . 34
3.4 Summary . . . 35

CHAPTER 4 PROBLEM STATEMENT, HYPOTHESIS, RESEARCH OBJECTIVES AND METHODOLOGY . . . 36
4.1 Problem Statement . . . 36
4.2 Hypothesis . . . 36
4.3 Objective . . . 36
4.4 Data Acquisition and Pre-processing . . . 37
4.4.1 SynthRad2023 brain dataset . . . 37
4.4.2 HNC dataset . . . 37
4.5 Methodology Overview . . . 38
4.5.1 Stage 1: MRI-to-CT synthesis . . . 38
4.5.2 Stage 2: CT-to-CBCT deformable Registration . . . 41
4.5.3 MRI/sCT-to-CBCT Registration Evaluation During Testing . . . 42

CHAPTER 5 ARTICLE 1: DIFFUSION-BASED IMAGE SYNTHESIS FOR DEFORMABLE MRI TO CBCT REGISTRATION WITH CROSS-ATTENTION IN HEAD AND NECK RADIOTHERAPY . . . 43
5.1 Introduction . . . 44
5.2 Materials and Methods . . . 47
5.2.1 Datasets and preprocessing . . . 47
5.2.2 Affine pre-alignment for MRI–CT and CT–CBCT pairs . . . 49
5.2.3 Image synthesis with MC-IDDPM . . . 50
5.2.4 Deformable registration with XMorpher . . . 52
5.2.5 Comparative methods . . . 53
5.2.6 Implementation details . . . 55
5.3 Performance evaluation . . . 56
5.3.1 sCT evaluation . . . 56
5.3.2 CT to CBCT registration evaluation . . . 57
5.3.3 sCT to CBCT registration evaluation using test set . . . 58
5.4 Results . . . 58

5.4.1 MRI-to-CT synthesis . . . . . 58
5.4.2 sCT-to-CBCT registration result . . . . . 64
5.4.3 MRI-to-CT synthesis and sCT-to-CBCT registration results . . . . . 65
5.5 Discussion . . . . . 65
5.6 Conclusion . . . . . 68

CHAPTER 6 GENERAL DISCUSSION . . . . . 70
6.1 Summary of the Synthesis Stage . . . . . 70
6.1.1 Limitations of the Synthesis Stage . . . . . 71
6.2 Summary of Deformable Registration Stage . . . . . 72
6.2.1 Evaluating Deformation Smoothness and Statistical Tests . . . . . 73
6.2.2 Limitations of the Registration Stage . . . . . 73
6.3 Summary of the HNC dataset Pre-Processing and limitations . . . . . 74
6.4 Future work . . . . . 75

CHAPTER 7 CONCLUSION . . . . . 76

REFERENCES . . . . . 78

# LIST OF TABLES

Table 5.1 MRI-to-CT synthesis image similarity metrics on the SynthRad2023 dataset using 5-fold cross-validation. . . . . 61

Table 5.2 MRI-to-CT synthesis image similarity metrics on the HNC dataset using 5-fold cross-validation comparing MC-IDDPM with CGAN and FCGAN. . . . . 61

Table 5.3 Dose metrics comparing predicted doses from sCTs on the HNC dataset. TTA denotes test-time augmentation. . . . . 63

Table 5.4 Registration performance on the HNC dataset using 5-fold cross-validation. Metrics reported include NCC (normalized cross-correlation), SSIM (structural similarity index measure), and TRE (target registration error in mm). . . . . 65

# LIST OF FIGURES

Figure 2.1 Sagittal illustration of head and neck anatomical structures. The image is taken from [2]. . . . . 4

Figure 2.2 (A) Patient positioning inside the scanner with the main static magnetic field ($B_0$) oriented along the head–foot axis; in diagrams, the z-axis is conventionally shown as vertical. (B) Illustration of proton precession: in the presence of $B_0$, hydrogen nuclei spin about their own axis while simultaneously precessing around the field direction. Images A and B are taken from [3]. (C) shows an axial slice example of an MRI from the HNC dataset. . . . . 6

Figure 2.3 Left: X-rays emitted from the source pass through the object and are captured by the X-ray detector. As the object rotates, the detector records varying transmission patterns, producing a sequence of projection images. Taken from [4]. Right: shows an axial slice example of a CT from the HNC dataset. . . . . 7

Figure 2.4 Illustration of the CBCT acquisition setup (left) and an axial slice example of a CBCT from the HNC dataset (right). . . . . 8

Figure 2.5 Columns from left to right: sagittal view of a pCT slice from one HNC patient with GTV (red) and OARs (various colors), radiation dose distribution with the same delineations, and the corresponding DVH. 10

Figure 2.6 Overview of the workflow for doing ART in HNC dataset. First, the delineation of PTVs and OARs is performed on the planning CT, and the corresponding dosimetry plan is then generated. During the course of treatment, patients may undergo anatomical changes, as illustrated on the right side of the figure ($\text{CBCT}_0$ taken the first day of treatment vs. $\text{CBCT}_{30}$ taken the 30th day of treatment). The objective of ART is to deformably register the pCT, along with its associated segmentations and dose plan, to the current patient anatomy to ensure more accurate treatment delivery. . . . . 11

Figure 3.1 A multilayer neural network. Each hidden layer applies a linear transformation (with weights $\mathbf{W}^{(k)}$ and biases $\mathbf{b}^{(k)}$), followed by a nonlinear activation function. A bias node with a constant input of 1 is included at each layer. . . . . 13

Figure 3.2 Illustration of a CNN architecture for image classification. The network consists of successive convolutional and pooling layers for feature extraction, followed by a flattening operation and fully connected layers for classification. The final SoftMax activation produces a probabilistic distribution over output classes. The image is taken from [5]. . . . . 14

Figure 3.3 U-Net architecture. Blue boxes indicate multi-channel feature maps, with their channel counts shown on top and spatial dimensions at the lower left. Arrows denote operations such as convolution, max pooling, and up-convolution. White boxes represent copied feature maps used in skip connections. The image is taken from [6]. . . . . 15

Figure 3.4 Training of a CGAN, in which the generator G learns to map an edge (condition y) to a photo and the discriminator D learns to distinguish between fake (synthesized/$G(y)$) and real photo. In contrast to an unconditional GAN, both the generator and the discriminator receive the input edge map as a condition. The image is taken from [7]. . . . 19

Figure 3.5 Overview of the VoxelMorph pipeline. The network learns parameters for a function $g$ that registers a moving volume $M$ to a fixed volume $F$. During training, the deformation field $\phi$ is used to warp $M$ via a spatial transformer. The loss function compares $M(\phi)$ with $F$ and promotes smoothness in $\phi$. The image is taken from [8]. . . . . 30

Figure 3.6 Overview of the PC-Reg-RT framework. The registration task is decomposed into two stages: perception and correspondence. The perception CNN extracts anatomical ROIs with clear boundaries, while the correspondence CNN performs accurate texture-preserving registration. Reverse Teaching enriches the perception CNN with structural and style knowledge from unlabeled images, enabling few-shot learning. The image is taken from [9]. . . . . 31

Figure 3.7 Overview of the XMorpher architecture. Dual U-Net-like transformer branches extract and fuse features from fixed and moving images using window-based cross-attention mechanisms, producing a deformation field $\phi$ optimized via similarity and smoothness losses. The image is taken from [10]. . . . . 33

Figure 4.1 HNC pre-processing. The figure shows the affine and rigid registration steps used in preprocessing, indicated by blue and green arrows, respectively. First, the CT is affine-registered to $\text{CBCT}_0$, then the MRI is affine-registered to the aligned CT, and finally, all other CBCTs are rigidly registered to $\text{CBCT}_0$. . . . . 38

Figure 5.1 Training overview for two stages: *(1) Synthesis* (MC-IDDPM) and *(2) Registration* (VM-XMorpher). In synthesis training stage (Green box above), given MRI, diffusion timestep $n$, and a noisy CT ($x_n$), the model (Swin V-Net denoiser) predicts the noise $\epsilon_\theta$ and the variance coefficient $k_\theta$ and is optimized with $\text{MAE}(\epsilon_\theta, \epsilon_n) + \gamma L_{\text{var}}$. The forward diffusion from $x_0$(clean CT) to $x_n$, in which a clean CT is added (represented by "+") with noise from a Gaussian distribution, is illustrative and not part of the training pipeline; $x_{n-1}$ is shown only to indicate the reverse update and is not used during training. The "Timestep" box represents the sinusoidal embedding that converts the current diffusion step n into a 128-dimensional vector. This embedding is then expanded by a linear (timestep-expanding) layer before being injected into the layers of the Swin V-Net denoiser, allowing each block to know the exact diffusion step and adjust its denoising accordingly. In registration training stage (blue box below), VM-XMorpher, which uses a Cross-Attention Transformer (CAT) blocks, learns a deformation field $\phi$ from CT/CBCT pairs using a local NCC loss (window = 9) with smoothness regularization; the spatial transformer network (STN) applies $\phi$ to the moving CT within the training pipeline. . . . . 48

Figure 5.2 Overview of the testing (inference) pipeline: the process performs end-to-end MRI-to-CBCT registration, where MC-IDDPM first synthesizes a CT from the input MRI, and VM-XMorpher subsequently aligns it with the daily CBCT. VM-XMorpher predicts a deformation vector field $\phi$ from the $(\text{sCT}, \text{CBCT})$ pair. The spatial transformer network (STN) then applies $\phi$ to the sCT and MRI to generate the warped images. . . . . 49

Figure 5.3 SynthRad2023 brain synthesized CT (sCT) comparison. sCT images were generated from the SynthRad2023 Task 1 brain validation set. The first row displays ground-truth CT images (left) and the corresponding input MRIs (right). Rows two to four show the sCT outputs from CGAN (row 2), FCGAN (row 3), and MC-IDDPM (row 4). Columns one, three, and five present the sCTs, while columns two, four, and six show the corresponding difference maps relative to the ground-truth CTs. . . . . 59

Figure 5.4 MRI-to-sCT and dose comparison using HNC validation set. The first row shows the input MRI and the synthesized CT (sCT) images generated by CGAN, FCGAN, and MC-IDDPM. The second row presents the corresponding ground-truth CT and the difference map for each method. The third row displays the real dose distribution computed on the ground-truth CT alongside the dose maps calculated on each sCT. The last row shows the dose-difference maps for the same models. 60

Figure 5.5 Dice similarity coefficient (DSC) values for selected anatomical structures using CGAN, FCGAN, and MC-IDDPM. Ground-truth RT segmentations (GTseg) are compared with TotalSegmentator (TotalSeg) predictions on sCT. A red dashed line indicates the DSC of spinal cord segmentations between GT and TotalSeg on real CT (rCT) for the spinal cord. . . . . 62

Figure 5.6 Boxplots of DVH differences for various anatomical structures using three synthesis models (CGAN, FCGAN, and MC-IDDPM), with and without test-time augmentation (TTA). Each box shows the distribution across five-fold cross-validation; structures include organs at risk (OARs) and planning target volumes (PTVs). . . . . 63

Figure 5.7 VM-XMorpher MRI-to-CBCT registration on test set. Each row shows: (1) Input moving MRI, (2) The generated moving sCT from the MRI and its segmentations, (3) moved sCT and segmentations, (4) fixed CBCT with DVF, (5) moved MRI and segmentations, (6) fixed CBCT with deformed segmentations, and (7) the 3D deformed sCT and DVF. Moving sCT segmentations are radiotherapy reference segmentations that are deformed to the sCT using affine and B-spline transformations. The listed segmentations are those deformed, and not all of them appear in this visualization slice. Cilinical target volume (CTV) 3, CTV 2, and CTV 1 correspond to PTV56, PTV63, and PTV70, respectively. 66

# LIST OF SYMBOLS AND ACRONYMS

| | |
|---|---|
| HNC | Head and Neck cancer |
| ART | Adaptive Radiotherapy |
| OAR | Organ At Risk |
| CT | Computed tomography |
| CBCT | Cone-beam computed tomography |
| MRI | Magnetic Resonance Imaging |
| DVF | Deformation vector field |
| DL | Deep Learning |
| HU | Hounsfield Units |
| CNN | Convolutional neural network |
| sCT | synthesized CT |
| CGAN | Conditional Generative Adversarial Network |
| FCGAN | Feature-consistency Conditional Generative Adversarial Network |
| DDPM | Denoising Diffusion Probabilistic Model |
| MC-IDDPM | MRO-to-CT Improved Denoising Diffusion Probabilistic Model |

# CHAPTER 1 INTRODUCTION

According to the World Health Organization, cancer is the second leading cause of death worldwide, responsible for approximately 9.6 million deaths each year, about one in every six deaths. Head and neck cancer (HNC) is among the most common cancer types globally and represents a substantial contributor to the global cancer burden. Its primary risk factors are tobacco and alcohol use, and infection with human papillomavirus (HPV). Treatment for HNC typically involves surgery, chemotherapy, or radiotherapy. Managing HNC with radiotherapy is particularly challenging due to the complex anatomy of the head and neck region and the proximity of numerous organs at risk (OARs), where unintended radiation damage can result in severe functional impairment and reduced quality of life during and after treatment. Radiotherapy, one of the most widely used treatments for HNC, therefore demands highly precise planning to ensure adequate tumor coverage while minimizing radiation exposure to surrounding healthy tissues. In this context, adaptive radiotherapy (ART), a promising approach, offers the opportunity to improve treatment outcomes by dynamically modifying treatment plans in response to daily anatomical variations.

In clinical practice, multiple imaging modalities are employed to support diagnosis, treatment planning, and monitoring of HNC. Medical imaging for patients with HNC typically involves Magnetic Resonance Imaging (MRI) for tumor detection, diagnosis and treatment plan. Computed tomography (CT) is acquired prior to treatment to facilitate tumor delineation, identify OARs, and develop a dose plan, as it provides the electron density information necessary for accurate dose calculation. Cone-beam computed tomography (CBCT) is acquired daily before each treatment session to monitor anatomical changes and assess tumor and treatment response over time. However, CBCT lacks electron density information, making it unsuitable for direct dose calculation. It is preferred over repeated CT scans due to its faster acquisition time and reduced radiation dose. The challenge arises because the dose plan, created on the planning CT (pCT), is not aligned with the daily CBCT due to anatomical changes in the patient during the treatment course. This anatomical variation necessitates accurate alignment between high-quality CT images and lower-quality CBCT images, which belong to different modalities and thus present challenges. Compared to CT, CBCT generally exhibits increased noise, reduced soft-tissue contrast, and various acquisition artifacts, further complicating the registration process.

Although MRI is non-invasive and provides more accurate tumor delineation, CT remains indispensable for dose planning. An MRI-based framework that enables ART is a promising

approach that could reduce both imaging time and patient exposure to radiation by eliminating the need to acquire a separate pCT scan. Achieving this requires reliable cross-modality image translation techniques to bridge the gap between MRI and CT.

With the rapid advancements in machine learning, particularly Deep Learning (DL), synthetic image generation has emerged as a promising approach for an MRI-based framework that enables ART. These techniques enable the generation of synthetic CT (sCT) images from MRI, which can then be used for downstream tasks such as segmentation or dose calculation. Furthermore, recent advances in deformable medical image registration enable more accurate alignment of the pCT or even sCT, along with its associated segmentations and dosimetric information, to daily CBCT acquisitions. This facilitates adaptation of the treatment plan to account for the patient's anatomical changes over the course of therapy, thereby enhancing treatment precision.

The objective of this project is to leverage and develop DL synthesis and registration techniques to enhance MRI-based approaches to support ART. The first part of this study focuses on synthesizing CT images from MRI, with the aim of generating sCTs that can potentially be used for dose calculation and segmentation. The second part involves applying deformable medical image registration to align CTs or corresponding sCT images and MRI images with their daily CBCT scans. The novelty of this work is to jointly evaluate diffusion-based MRI-to-CT synthesis and transformer-based CT-to-CBCT deformable registration on the HNC dataset. By demonstrating that MRI-derived sCTs can support segmentation, dose estimation, and deformable alignment, the proposed pipeline moves toward an MRI-only ART workflow that minimizes dependence on planning CT while maintaining clinical accuracy.

## 1.1 Outline of the thesis

Chapter 2 introduces the clinical background of HNC treatment and the principles of ART, followed by an overview of the imaging modalities used in this study and their respective roles within the proposed workflow.

Chapter 3 reviews DL methodologies and relevant literature on medical image synthesis and deformable registration, establishing the theoretical and technical foundations, identifying their limitations, and justifying the selection of the proposed approach in the fourth chapter.

Chapter 4 details the methodology, describing the datasets, preprocessing steps, new model architectures, and training strategies.

Chapter 5 presents the research paper resulting from this master's study, describing the complete two-stage pipeline.

Chapter 6 summarizes the work and provides an interpretation of the results presented in the paper, along with a general discussion of the study's limitations and directions for future work.

Chapter 7 concludes the thesis by summarizing the findings from the proposed MRI-based ART pipeline and discussing its clinical implications.

# CHAPTER 2 BACKGROUND

## 2.1 Head and Neck Cancer

Head and neck cancer (HNC) refers to a group of cancers that develop in areas such as the upper portions of the respiratory and digestive tracts and the oral region. The main risk factors include alcohol and tobacco consumption and infection with certain types of human papillomavirus. These cancers are typically diagnosed using imaging techniques like CT and MRI, along with a biopsy to confirm the cancer type. Once diagnosed, treatment depends on the location and stage of the cancer and may involve surgery, radiation therapy, or chemotherapy [11]. This study focuses on treatment with radiotherapy.

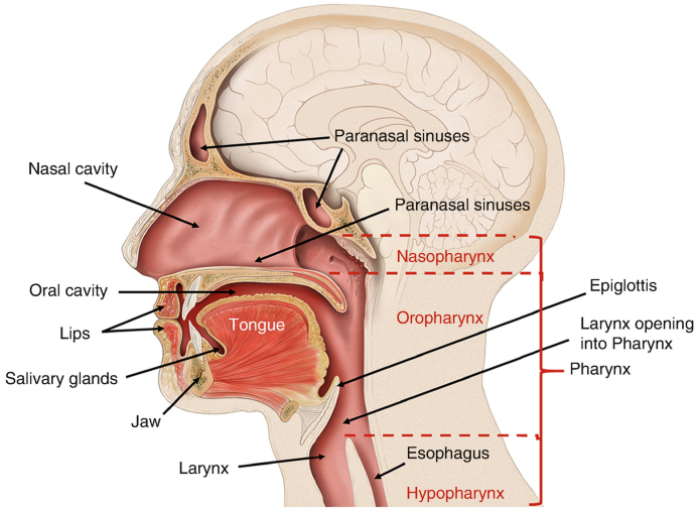


Figure 2.1 Sagittal illustration of head and neck anatomical structures. The image is taken from [2].

### 2.1.1 Anatomical Structures in HNC

Radiation therapy for HNC must account for the proximity of critical anatomical structures that are highly sensitive to radiation. These organs at risk (OARs) include the salivary glands, optic nerves, auditory system, voice and swallowing structures, and the brain. Damage to these structures can lead to severe side effects, such as dysphagia (difficulty swallowing), dysphonia (impaired speech), or respiratory distress, which significantly impact patient quality of life. Therefore, maximal preservation of OARs is a central objective in radiotherapy treatment [11].

### 2.1.2 Adaptive Radiotherapy in HNC

Given the critical importance of sparing OARs, the primary objective of radiotherapy is not only to deliver the prescribed radiation dose precisely to the tumor, but also to minimize irradiation of surrounding healthy tissues. In adaptive radiotherapy (ART), the treatment plan is dynamically updated throughout the treatment course to account for daily anatomical changes, ensuring consistent target coverage while further reducing the dose delivered to nearby OARs. After the tumor is diagnosed using MRI (or CT), radiation therapy planning is conducted by dosimetrists and radiation oncologists, beginning with the contouring of three hierarchical target volumes on CT: the Gross Tumor Volume (GTV), representing the visible tumor. The Clinical Target Volume (CTV), which includes the GTV plus areas suspected of microscopic disease, and the Planning Target Volume (PTV), which adds margins to account for setup errors and internal motion. The prescribed radiation dose for the CTV typically ranges from 66 to 70 Gray (Gy), delivered in daily fractions of 2 Gy. This dose is not delivered all at once, but rather divided into daily sessions (called fractions), with each session delivering 2 Gy of radiation. As a result, the full course of treatment spans approximately 33 to 35 sessions. Meanwhile, OARs are contoured and assigned specific dose constraints to limit radiation-induced toxicity [11].

## 2.2 Medical Imaging for HNC

Medical imaging plays a critical role in the management of HNC, guiding tumor detection, radiotherapy planning, and treatment monitoring. The modalities used are MRI, CT, and CBCT, each serving complementary purposes. Dose planning is also an essential modality involved in the treatment process.

MRI provides superior soft tissue contrast and is used for tumor diagnosis, severity assessment, and treatment planning. CT is essential for radiotherapy planning because it provides electron density information required for accurate dose calculations. Additionally, the delineation of the tumor and OARs is performed on the planning CT (pCT).

During the several weeks of radiotherapy, patients often experience anatomical changes, such as weight loss and tumor shrinkage, which can compromise treatment accuracy. While CT could theoretically be used to track these changes, its repeated use would expose the patient to unnecessary radiation. Instead, CBCT is routinely employed for daily imaging during treatment, as it can be acquired within a minute, whereas CT requires significantly more time. Consequently, CBCT facilitates frequent monitoring of anatomical changes and verification of patient setup, supporting ART without introducing a significant additional radiation burden

[12].

A key challenge in this workflow is accurately registering the pCT, its associated segmentations (tumor and OARs), and the dose plan with the daily CBCT scans. This difficulty arises from CBCT's inherent limitations: low spatial resolution that obscures fine anatomical details, inconsistent Hounsfield units that hinder reliable density mapping, and reconstruction artifacts that complicate deformable registration of the tumor and OARs, particularly when patient anatomy changes during treatment.

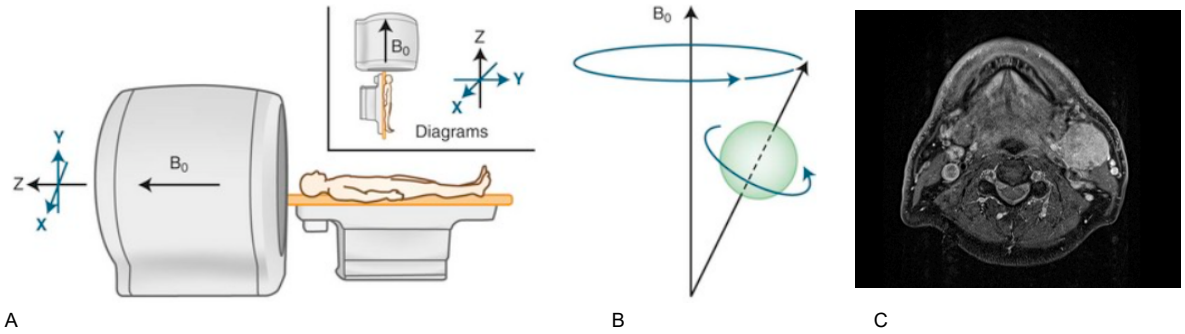


Figure 2.2 (A) Patient positioning inside the scanner with the main static magnetic field ($B_0$) oriented along the head–foot axis; in diagrams, the z-axis is conventionally shown as vertical. (B) Illustration of proton precession: in the presence of $B_0$, hydrogen nuclei spin about their own axis while simultaneously precessing around the field direction. Images A and B are taken from [3]. (C) shows an axial slice example of an MRI from the HNC dataset.

### 2.2.1 Magnetic Resonance Imaging (T1-weighted)

MRI excels at visualizing soft tissues, such as tumor boundaries, and in the HNC dataset, it is primarily used for diagnosis, staging, and assessment of disease severity. With its superior soft-tissue contrast compared to CT, MRI enables more reliable identification of tumor extent, characterization of disease, and confirmation of severity and type. Although target and OAR delineation for radiotherapy planning is performed on CT, MRI provides complementary information that guides interpretation and supports the design of more accurate treatment plans.

MRI (T1-weighted) is a non-invasive technique that generates detailed anatomical images by detecting signals from hydrogen nuclei (protons) in water and fat. When placed in a strong static magnetic field ($B_0$), these protons align. A brief radiofrequency pulse disrupts this alignment. After the pulse ceases, the protons release energy as they recover their alignment

with $B_0$ along the longitudinal axis. The characteristic time constant for this recovery process is the T1 relaxation time, which varies between different biological tissues due to their molecular environment. T1-weighted sequences are specifically designed to produce image contrast based on these inherent T1 differences. This is achieved by using a short Repetition Time (TR), preventing full recovery of longitudinal magnetization between pulses and leaving tissues with different T1 values at distinct recovery stages. A short Echo Time is also used to minimize signal contributions from other relaxation mechanisms. Consequently, tissues with shorter T1 values (such as fat) recover more longitudinal magnetization during the brief TR interval, yielding a stronger signal and appearing bright in the final image. This contrast makes T1-weighted MRI particularly valuable for structural analysis [13]. T1-weighted MRI from the HNC dataset was employed in this study.

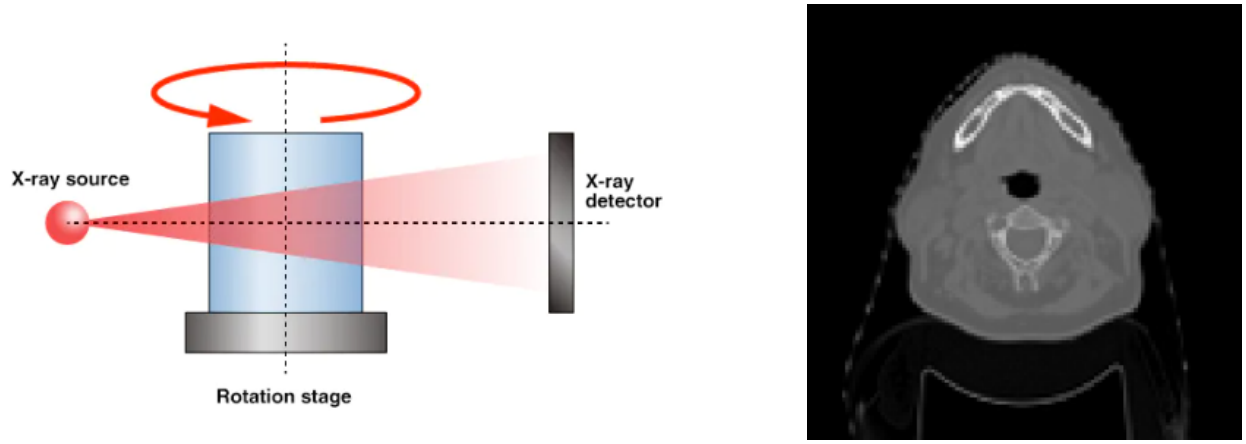


Figure 2.3 Left: X-rays emitted from the source pass through the object and are captured by the X-ray detector. As the object rotates, the detector records varying transmission patterns, producing a sequence of projection images. Taken from [4]. Right: shows an axial slice example of a CT from the HNC dataset.

### 2.2.2 Computed Tomography

CT is a prevalent clinical imaging technique that delivers high-resolution anatomical details at a lower cost than modalities like MRI; however, unlike MRI, it is not non-invasive. CT generates cross-sectional images by capturing X-ray projections from multiple angles around the patient. Computational reconstruction of these projections produces slices composed of pixels, while stacked slices form a three-dimensional volume of voxels. The slice thickness is adjustable to meet spatial resolution requirements

Each voxel encodes the attenuation of X-rays by biological tissues, determined by their density

and atomic number. These values are standardized in Hounsfield Units (HU), calculated as:

$$\mathrm{HU} = 100 \times \frac{\mu_{\mathrm{tissue}} - \mu_{\mathrm{water}}}{\mu_{\mathrm{water}}} \tag{2.1}$$

Where $\mu$ denotes the attenuation coefficient. For instance, bones and other dense tissues exhibit higher, positive HU values, whereas less dense tissues, such as fat and air, have lower values.

In HNC radiotherapy planning, CT is the modality used for tumor and OAR delineation. [11].

### 2.2.3 Cone Beam Computed Tomography

CBCT is integral to modern image-guided radiation therapy (IGRT). The system rotates 360° around the patient, capturing rapid, low-dose 2D X-ray projections using a cone-shaped beam and flat-panel detector. These projections are reconstructed into a 3D volumetric dataset in approximately one minute, providing real-time anatomical visualization immediately prior to or during treatment delivery while reducing patient discomfort. Despite its utility, CBCT exhibits lower spatial resolution, diminished soft-tissue contrast, image noise, reconstruction artifacts, and inaccurate HU values compared to conventional CT. Consequently, unreliable HU values render CBCT unsuitable for primary diagnosis or direct dose calculation [11].

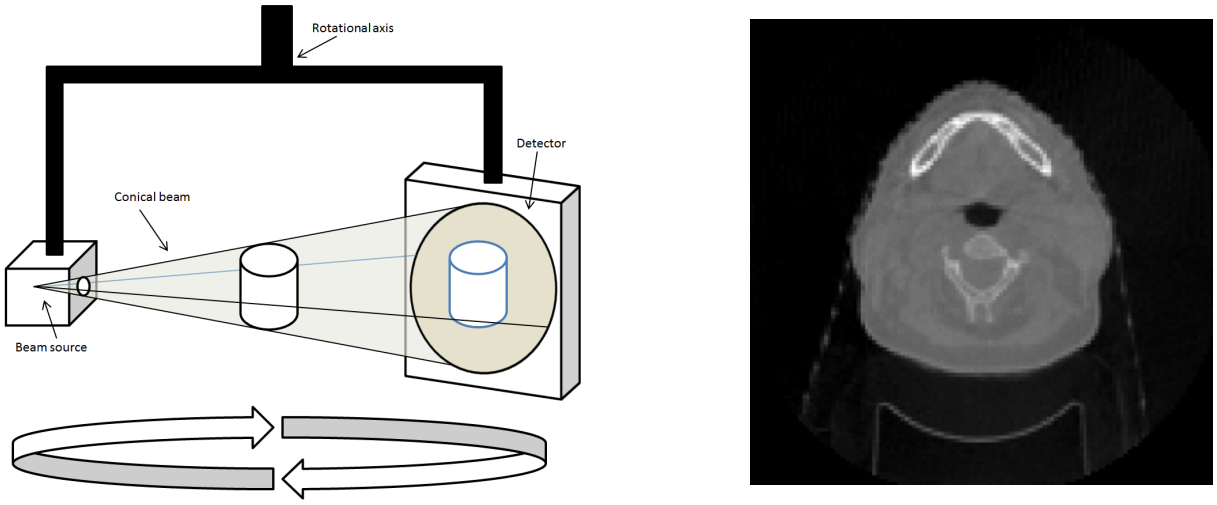


Figure 2.4 Illustration of the CBCT acquisition setup (left) and an axial slice example of a CBCT from the HNC dataset (right).

In head and neck radiotherapy, CBCT fulfills two primary functions:

**(1) Daily Setup Verification**

The intricate anatomy of the head and neck requires high precision. Pre-treatment CBCT scans enable clinicians to detect and correct translational and rotational positioning errors through rigid registration with the pCT. This process ensures accurate tumor targeting and sparing of OARs [11].

**(2) Adaptive Radiotherapy Support**

Significant anatomical changes (e.g., tumor shrinkage, weight loss) that occur over the typical 6–7 week treatment course can necessitate modifications to the planned dose distributions. Serial CBCT imaging monitors these changes longitudinally. While rigid registration corrects daily setup variations, substantial anatomical deformations often necessitate deformable image registration (DIR) to accurately map tissue changes and their dosimetric impact. This facilitates ART, allowing plan modification to maintain optimal tumor coverage and OAR protection [11].

### 2.2.4 Dose Planning

In radiotherapy for HNC, dose planning begins with the delineation of target volumes and OARs on the planning CT (pCT). Using these contours, optimization algorithms generate a three-dimensional dose distribution that conforms to the tumor while respecting dose constraints for surrounding structures. Modern delivery techniques allow highly conformal irradiation by modulating beam intensity and angles. The final plan, prescribed in daily fractions, is reviewed by the clinical team to ensure adequate tumor coverage and OAR sparing. During treatment, daily CBCT scans are used to verify alignment and guide adaptations when anatomical changes compromise the accuracy of the original plan [11].

Figure 2.5 illustrates pCT from the HNC dataset, with the GTV and OARs delineated on it, and radiation dose distribution where the tumor receives a high dose while surrounding OARs are spared. The figure also presents the Dose-Volume Histogram (DVH), a graphical representation that illustrates how the volume of a specific ROI is distributed across varying radiation dose levels. It provides a quantitative basis for evaluating treatment plans by determining whether target volumes receive adequate dose coverage and whether OARs are sufficiently spared.

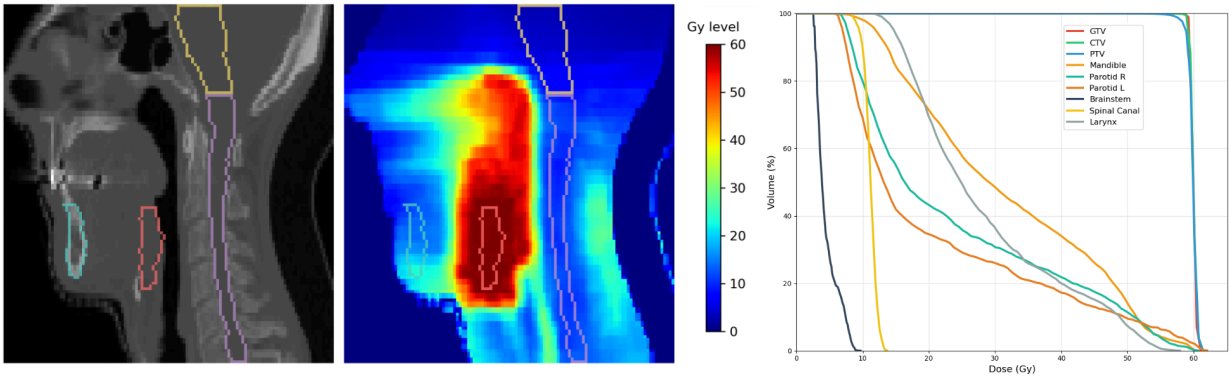


Figure 2.5 Columns from left to right: sagittal view of a pCT slice from one HNC patient with GTV (red) and OARs (various colors), radiation dose distribution with the same delineations, and the corresponding DVH.

## 2.3 The workflow for ART

In ART, where the goal is to adapt the treatment according to the patient's current anatomy, the pCT and its associated segmentations of the tumor, OARs, and dose plan are deformably registered to the daily CBCTs acquired throughout the treatment course. This process allows tracking of anatomical and positional variations that occur over time, such as tumor shrinkage, organ deformation, or patient setup changes. By propagating the contours and dose information from the pCT to each daily scan, clinicians can accurately monitor anatomical evolution and assess whether the delivered dose distribution remains consistent with the original treatment intent. When substantial deviations in anatomy or dose coverage are identified, the treatment plan can be adapted by aligning the dose, target volumes, and OARs to ensure adequate tumor coverage while minimizing unnecessary exposure to healthy OAR tissue. Consequently, ART enables a more precise and patient-specific approach to radiotherapy, maintaining treatment accuracy and safety even as the patient's anatomy changes. Figure 2.6 illustrates the ART workflow used in this study.

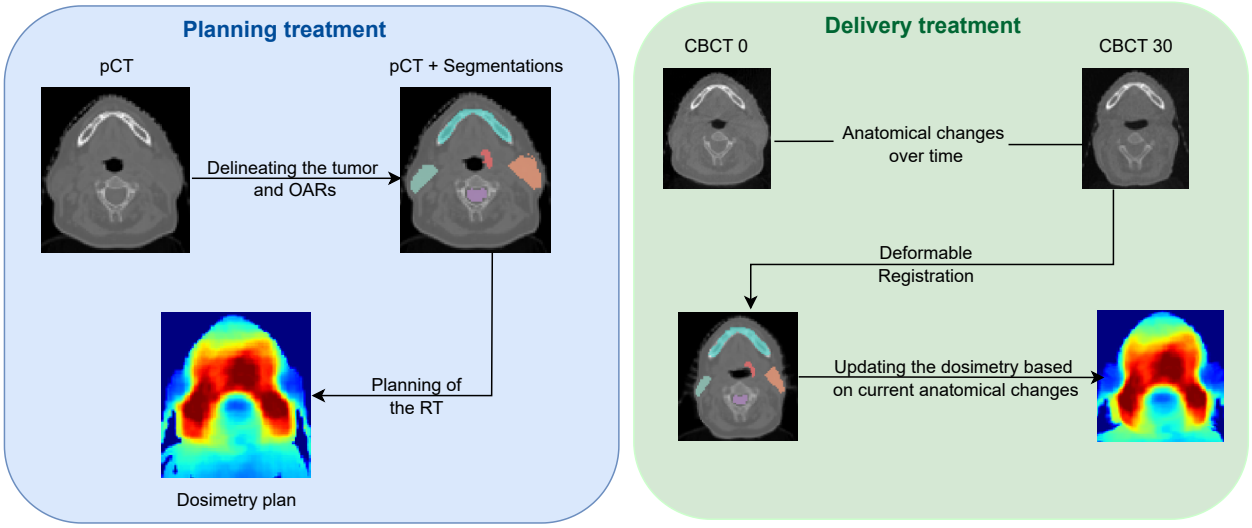


Figure 2.6 Overview of the workflow for doing ART in HNC dataset. First, the delineation of PTVs and OARs is performed on the planning CT, and the corresponding dosimetry plan is then generated. During the course of treatment, patients may undergo anatomical changes, as illustrated on the right side of the figure ($CBCT_0$ taken the first day of treatment vs. $CBCT_{30}$ taken the 30th day of treatment). The objective of ART is to deformably register the pCT, along with its associated segmentations and dose plan, to the current patient anatomy to ensure more accurate treatment delivery.

# CHAPTER 3 LITERATURE REVIEW

## 3.1 Deep Learning Concepts

### 3.1.1 Multilayer Neural Network

Multilayer neural networks learn hierarchical representations of data, transforming raw inputs into a new representation of the data (e.g., converting pixel values to edge detectors, then to shape recognizers) to solve complex tasks.

A multilayer neural network is a computational model composed of sequentially connected layers of neurons (Figure 3.1). Each layer receives input from the previous layer, applies a linear transformation, and passes the result through a nonlinear activation function (e.g., ReLU or sigmoid). This hierarchical structure enables the network to automatically learn new representations from data.

Let the input be $\mathbf{x}$, and suppose the network has $L$ hidden layers. Let $k$ denote the layer index, where $k > 0$:

$$\text{Pre-activation}: \quad \mathbf{a}^{(k)}(\mathbf{x}) = \mathbf{b}^{(k)} + \mathbf{W}^{(k)}\mathbf{h}^{(k-1)}(\mathbf{x}) \tag{3.1}$$

$$\text{Activation}: \quad \mathbf{h}^{(k)}(\mathbf{x}) = g\left(\mathbf{a}^{(k)}(\mathbf{x})\right) \tag{3.2}$$

$$\text{Output}: \quad f(\mathbf{x}) = \mathbf{h}^{(L+1)}(\mathbf{x}) = o\left(\mathbf{a}^{(L+1)}(\mathbf{x})\right) \tag{3.3}$$

Here, $g(\cdot)$ and $o(\cdot)$ are element-wise nonlinear activation functions, and $f(\mathbf{x})$ denotes the final output of the network.

Loss and learning: To train the network, a loss function $\mathcal{L}(f(\mathbf{x}), y)$ is used to measure the discrepancy between the predicted output $f(\mathbf{x})$ and the target $y$. The parameters $\{\mathbf{W}^{(k)}, \mathbf{b}^{(k)}\}$ are optimized using gradient descent. For each layer, the updates are given by:

$$\mathbf{W}^{(k)} \leftarrow \mathbf{W}^{(k)} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{W}^{(k)}}, \quad \mathbf{b}^{(k)} \leftarrow \mathbf{b}^{(k)} - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{b}^{(k)}} \tag{3.4}$$

Where $\alpha$ is the learning rate. These gradients are computed via backpropagation using the chain rule. Stochastic Gradient Descent (SGD) updates the model's parameters after every single training example. This makes each update very fast and can help avoid getting stuck in local minima, but it may also lead to unstable learning. Batch Gradient Descent computes

the gradient using the entire training dataset before performing a single update. While this provides a precise direction for the update, it is slow and computationally intensive for large datasets. Mini-Batch Gradient Descent, the most commonly used method, splits the data into smaller groups called mini-batches. The model updates its parameters after processing each mini-batch, offering a balance between the speed and stability of SGD and the accuracy of full-batch updates. It combines the advantages of both approaches, making it efficient and scalable for practical training [14].

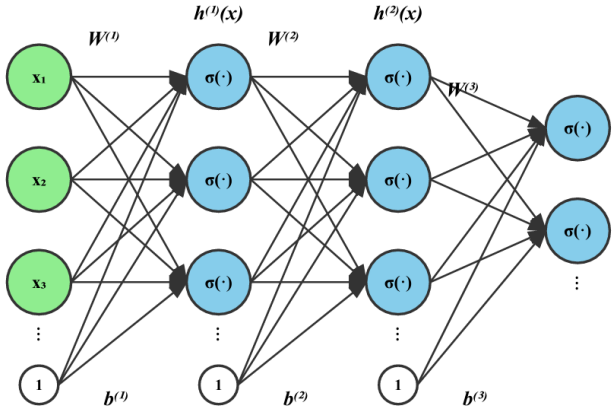


Figure 3.1 A multilayer neural network. Each hidden layer applies a linear transformation (with weights $\mathbf{W}^{(k)}$ and biases $\mathbf{b}^{(k)}$), followed by a nonlinear activation function. A bias node with a constant input of 1 is included at each layer.

### 3.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a specialized form of neural networks designed to handle image data more efficiently than traditional fully connected neural networks. One of their major advantages lies in how they manage spatial information through convolution and max pooling operations [15].

In CNNs, convolutional layers apply small filters (kernels) that slide over local regions of the input image, computing a weighted sum to detect features like edges or textures. This localized connectivity drastically reduces the number of parameters compared to traditional neural networks, which require every neuron to be connected to all pixels. As these layers are stacked, they enable the network to learn increasingly abstract features, from simple lines to complex shapes and objects [15].

After convolution, max pooling layers are typically applied. Max pooling performs downsam-

pling by selecting the maximum value from small regions (e.g., 2×2 blocks) of each feature map. This operation reduces the spatial resolution of the feature maps, meaning the image becomes smaller, but the most important features are preserved. At the same time, the receptive field, the portion of the input image each neuron "sees", increases in deeper layers, allowing the network to capture more global patterns [15].

Together, these mechanisms make CNNs both efficient and powerful. Unlike standard multilayer neural networks, which can overfit easily on high-dimensional image data due to their dense connections, CNNs exploit the structure of image inputs to drastically reduce parameters, prevent overfitting, and enhance generalization. Thus, CNNs achieve superior performance in image recognition tasks with fewer resources, making them the architecture of choice in modern computer vision [15].

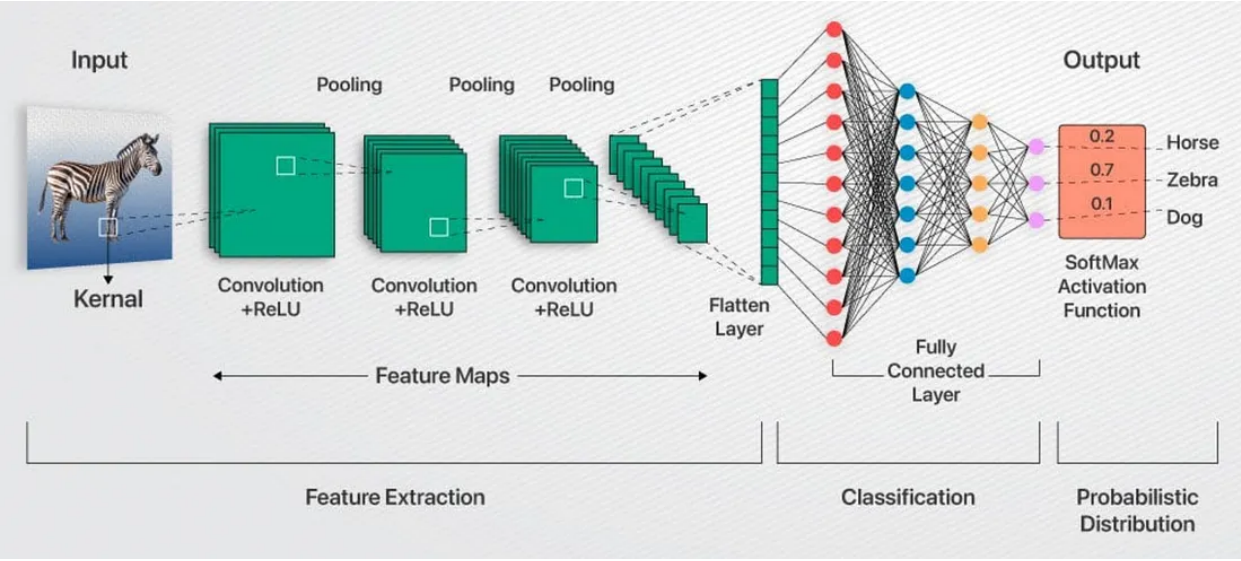


Figure 3.2 Illustration of a CNN architecture for image classification. The network consists of successive convolutional and pooling layers for feature extraction, followed by a flattening operation and fully connected layers for classification. The final SoftMax activation produces a probabilistic distribution over output classes. The image is taken from [5].

### 3.1.3 U-Net

U-Net [6] is a fully convolutional neural network that adopts a symmetric encoder-decoder structure (U-shaped). The left side of U-Net, shown in Figure 3.3, serves as the contracting path, which captures context through repeated use of two $3 \times 3$ convolutions, each followed by a ReLU activation and a $2 \times 2$ max-pooling operation for downsampling. The number of feature channels doubles at each step of downsampling [6].

The right side of the architecture performs upsampling using transposed convolutions (up-convolutions), reducing the number of channels by half. Importantly, each upsampled feature map is concatenated with the corresponding feature map from the contracting path. These skip connections help recover spatial resolution lost during downsampling by reintegrating high-resolution features from earlier layers [6].

As illustrated by the horizontal arrows in Figure 3.3, skip connections play a key role in combining semantic context with precise localization. After concatenation, two $3 \times 3$ convolutions (with ReLU) refine the merged features. The final $1 \times 1$ convolution produces the output [6]. U-Net is used as the backbone in many models.

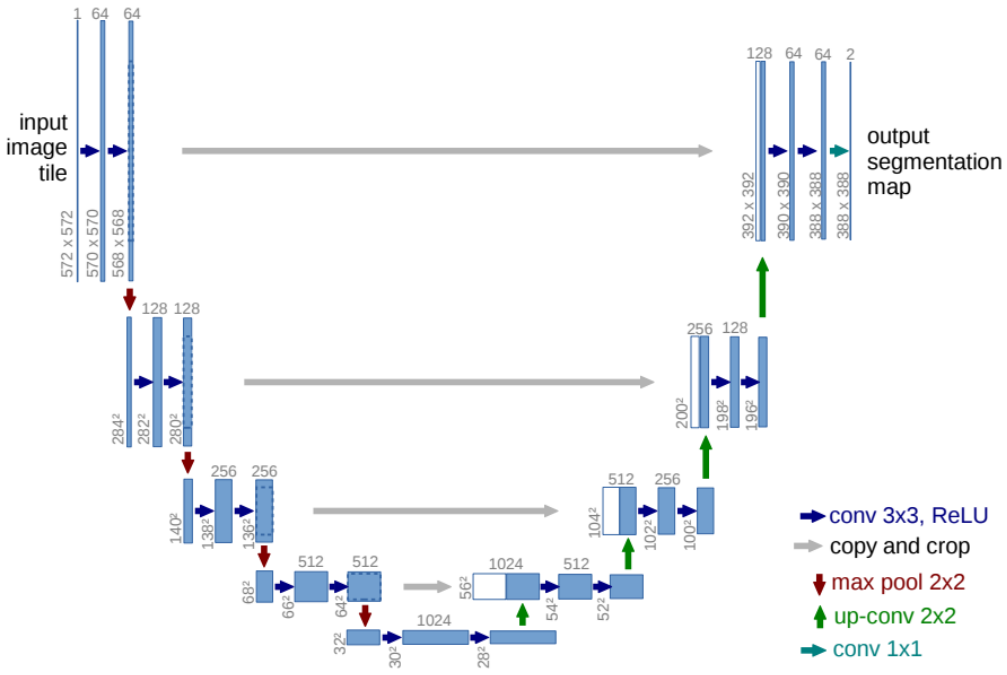


Figure 3.3 U-Net architecture. Blue boxes indicate multi-channel feature maps, with their channel counts shown on top and spatial dimensions at the lower left. Arrows denote operations such as convolution, max pooling, and up-convolution. White boxes represent copied feature maps used in skip connections. The image is taken from [6].

## 3.2 Medical Image Synthesis

Both MRI and CT are essential imaging modalities for diagnosis and for planning radiation therapy, as mentioned earlier. MRI provides superior soft tissue contrast [16–18], whereas CT is indispensable for accurate radiation dose calculation [19–21]. Generating CT images from MRI offers a promising approach to combine the strengths of both modalities, enhancing

patient comfort while reducing cost, acquisition time, and radiation exposure [22], thereby eliminating the need for a separate CT scan.

There are different methods for synthesizing CT from MRI images, and they are generally divided into three groups.

### 3.2.1 Classical Methods

#### Density-based

Morphological operations and predefined intensity thresholds are utilized to divide the image based on the image density into different class labels. Then a specific density value is assigned to each tissue region or class label, which is used to create the synthetic CT image [23, 24]. The threshold is needed to be adjusted manually, which is a problem with this method.

#### Atlas-based

In this method, one or more atlases of MRI with corresponding CT are registered to the patient's MRI. After registration, the CT information from these atlases is used to create labels for different organs, enabling the patient's MRI scan to be converted into a synthetic CT image [25] cite ghalate. The issue with this method is that it is sensitive to structural similarities between atlases and patient anatomy. Moreover, a suitable number of atlases are chosen manually, which can introduce errors when applied across different patients [26] [27].

Classical methods such as density-based and atlas-based techniques rely heavily on manual thresholds [24] and registration accuracy [27], which limit their adaptability across diverse anatomies.

### 3.2.2 Deep Learning Methods

To overcome the limitations of the traditional method, machine learning approaches, especially deep learning methods, have been developed that leverage high computational resources to automatically learn meaningful anatomical representations suited for the specific application, while accounting for differences across diverse anatomies [28]. Methods like CNNs [29] and Generative Adversarial Networks (GANs) [28] learn complex, non-linear relationships for image synthesis.

### Generative Adversarial Networks

GANs [28] are suitable for domain translation problems. They revolutionized image synthesis by training two neural networks, a generator and a discriminator, in an adversarial framework. The generator learns to produce realistic data distributions, while the discriminator distinguishes between real and synthetic samples. This paradigm has become foundational for tasks like image-to-image translation [28]. More specifically, the generator takes z as input, which is sampled from a uniform or Gaussian distribution. Then, the generator learns to map z from a low-dimensional space to a high-dimensional or real image space. The discriminator takes the images generated by the generator G(z) (or the real image) as input and classifies them as real or fake. The generator tries to generate images that look as real as possible so that the discriminator cannot distinguish them as fake. During training, the generator and discriminator improve by optimizing the objective function below.

$$\mathcal{L}_{\text{GAN}} = \mathbb{E}_{x \sim p(x)} \left[ \log D(x) \right] + \mathbb{E}_{z \sim p(z)} \left[ \log \left( 1 - D(G(z)) \right) \right] \tag{3.5}$$

$x \sim p(x)$ denotes that the data is from the real distribution, while $z \sim p(z)$ denotes that the data is from a uniform or Gaussian distribution(fake distribution). The discriminator tries to maximize the above objective function, while the generator tries to minimize it. $\mathbb{E}_X$ and $\mathbb{E}_Z$ denote expectation over x and z, respectively.

### Conditional GANs (CGAN)

The main limitation of an unconditioned generative model is its inability to control the mode of data generation. CGAN addresses this by conditioning both the generator and discriminator on auxiliary information, allowing for targeted and more controllable data synthesis [30]. CGAN is an extension of the original GAN that uses extra information y, such as class labels or data from another modality. In this setup, both the generator and discriminator are conditioned on y, allowing the model to guide the generation process based on specific attributes. The generator receives a combination of the noise vector z and the condition y, and generates an image based on this input, denoted as $p(z|y)$. Similarly, the discriminator takes both the image (real or generated) and the condition y as input. It learns to distinguish between real and fake image (condition pairs), which helps the generator improve over time [30].

The objective function of the CGAN is defined as a two-player minimax game:

$$\mathcal{L}_{\text{CGAN}} = \mathbb{E}_{x\sim p(x)}\left[\log D(x|y)\right] + \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D(G(z|y))\right)\right] \tag{3.6}$$

Same as the usual GAN, in CGAN the discriminator tries to maximize the above objective function, while the generator tries to minimize it.

**Pix2Pix**

Pix2Pix [7] extends the CGAN framework to supervised image-to-image translation tasks where each input image has a corresponding output image. Unlike the original CGAN [30], which takes both a random noise vector $z$ and a conditioning variable $y$, Pix2Pix removes the noise input and instead conditions the generation solely on the input image. This reflects a deterministic setting, where the generator learns a direct mapping from input to output.

To enhance the quality and stability of the generated outputs, Pix2Pix combines the standard adversarial objective with a pixel-wise L1 reconstruction loss. While prior work [31] explored the addition of L2 loss to guide the generator toward the ground truth, Pix2Pix favors L1 loss as it reduces blurring and better preserves image details. The combined objective is defined as:

$$\mathcal{L}_{\text{pix2pix}} = \mathcal{L}_{\text{CGAN}} + \lambda\mathcal{L}_1(G) \tag{3.7}$$

where the adversarial loss follows the CGAN formulation, with the noise component removed. $\mathcal{L}_1$ is computed between the generated output and the ground truth image, which promotes similarity between the generated output and the ground truth image. $\lambda$ is a hyperparameter that balances the two terms and can be set up to 100 [7]. The L1 loss is computed as:

$$\mathcal{L}_1(G) = \mathbb{E}_{x,y}[\|x - G(y)\|_1] \tag{3.8}$$

Here, $y$ is the input condition, $x$ is the corresponding ground truth, and $G(y)$ is the generated output. The discriminator $D$ attempts to distinguish real from synthetic one, while the generator tries to both fool the discriminator and match the ground truth image under the L1 norm.

Pix2Pix also introduces key architectural choices to improve image quality. The generator adopts a U-Net structure [6], to retain spatial details, especially useful when input and output images share structural alignment. Furthermore, instead of a traditional discriminator that outputs a single real/fake score, pix2pix uses a PatchGAN discriminator, which outputs a

small patch-based prediction instead of a single score, encouraging the generator to produce locally realistic textures [7].

Together, the L1-guided loss, UNet generator, and PatchGAN discriminator make Pix2Pix a strong and general framework for diverse image translation problems, outperforming earlier CGAN-based methods in both visual quality and structural accuracy [7].

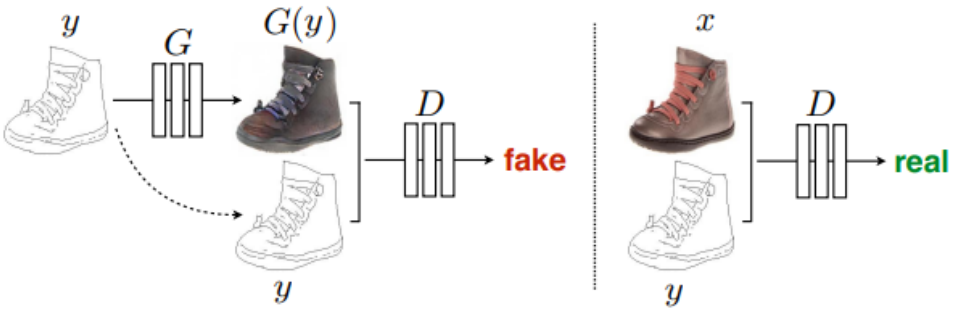


Figure 3.4 Training of a CGAN, in which the generator G learns to map an edge (condition y) to a photo and the discriminator D learns to distinguish between fake (synthesized/$G(y)$) and real photo. In contrast to an unconditional GAN, both the generator and the discriminator receive the input edge map as a condition. The image is taken from [7].

### GANs Limitations

GANs generally suffer from three major issues: mode collapse, convergence difficulties, and instability, all of which make training adversarial models more challenging.

**(1)Mode Collapse:** The generator produces low-diversity outputs by mapping distinct inputs to identical or nearly identical samples [32,33]. Deep Regret Analytic GAN (DRAGAN) mitigates this issue by penalizing the discriminator's gradients near real data [32,34].

**(2)Convergence Issues:** GAN training often fails to reach equilibrium due to the non-convex-concave nature of the optimization process, leading to oscillations or divergence in the generator and discriminator updates [28,33,35]. Spectral Normalization GAN (SN-GAN) addresses this by applying weight normalization to slow down the discriminator's learning, promoting more stable training dynamics [36].

**(3)Vanishing Gradients:** When the discriminator becomes too strong early in training, it provides near-zero gradients to the generator, hindering its ability to learn effectively [28, 33, 37]. To address this, [28] proposed using an alternative generator loss, $\mathcal{L}_G = -\log D(G(z))$ instead of the minimax objective. This non-saturating loss helps maintain

meaningful gradients for the generator, ensuring continued learning even when the discriminator performs well, thereby improving training stability [28].

GANs have additional limitations beyond the general challenges associated with generative adversarial networks. Although CGANs [38] leverage paired data to learn cross-modal mappings and have shown promising results for brain MRI-to-CT synthesis [39, 40], head and neck synthesis presents greater challenges due to anatomical complexity [41,42]. GAN-based approaches employing U-Net [6] generators have been applied to head and neck MRI-to-CT synthesis [41, 43, 44]; however, they typically rely on a combination of adversarial loss and pixel-wise losses such as L1 or L2 [43,45,46]. This often results in difficulty in preserving fine anatomical structures, a critical limitation [42] for radiotherapy dose calculation.

### Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPMs) have demonstrated superior sample quality to many GAN-based methods. On the unconditional CIFAR-10 dataset, DDPM [47] outperforms strong GAN baselines such as SNGAN [36] and SNGAN-DDLS [48], while avoiding the training instability and mode collapse issues characteristic of adversarial learning. Unlike GANs, diffusion models are trained without an adversarial objective, which contributes to their stability. DDPMs define a Markovian forward diffusion process that gradually adds Gaussian noise to real data, eventually converting it into pure Gaussian noise. Subsequently, a neural network learns the reverse diffusion process to iteratively remove this noise, reconstructing the original clean data from noisy inputs.

Mathematically, the forward diffusion process progressively transforms a real image into a noisy image over timesteps, where each step adds a small amount of Gaussian noise:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}) \tag{3.9}$$

where $x_0$ is the original data sample, $x_t$ is the noisy version at time step $t$, $\beta_t$ is the variance schedule, and $\mathbf{I}$ is the identity matrix. After many steps, the image becomes nearly pure noise: $x_T \sim \mathcal{N}(0, \mathbf{I})$. The reverse process is a learned denoising model that reconstructs $x_{t-1}$ from $x_t$. Instead of predicting $x_{t-1}$ directly, the model is trained to predict the noise added during the forward process. This reverse diffusion process, parameterized by a neural network, attempts to undo the noise addition by modeling the inverse conditional distribution:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3.10}$$

where $p_\theta$ is the learned reverse model, implemented via a neural network $\epsilon_\theta$, typically a U-Net [6], which predicts the noise $\epsilon$ added at each timestep.

$\mu_\theta(x_t, t)$ denotes the predicted mean of the reverse Gaussian distribution. $\Sigma_\theta(x_t, t)$ represents the variance of the reverse distribution; it is not learned but is instead assigned according to a predefined variance schedule $\beta_t$ or $\tilde{\beta}_t$, as in [47], which varies with the timestep $t$. $\tilde{\beta}_t$ is a function of $\beta_t$. $\theta$ represents the learnable parameters of the neural network. This process is repeated from $t = T$ down to $t = 1$ to progressively transform noise into a clean image sample.

During training, the network learns to predict the added noise at each timestep, typically optimized using the mean square error between the predicted and actual noise (This formulation corresponds to the simplified training objective proposed in [47]):

$$\mathcal{L}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon},\, t\right)\right\|^2\right] \tag{3.11}$$

$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ is the true noise, and $\epsilon_\theta$ is the model's predicted noise. $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t}(1 - \beta_s)$

The noisy image at time $t$ is sampled as:

$$x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1-\bar{\alpha}_t} \cdot \epsilon \tag{3.12}$$

To generate a new less noisy image, start from Gaussian noise $x_T \sim \mathcal{N}(0, \mathbf{I})$ and apply the learned reverse steps iteratively:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(\mathbf{x}_t, t)\right) + \sigma_t \mathbf{z}, \tag{3.13}$$

where $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise, and $\sigma_t^2 = \tilde{\beta}_t$ is the variance used during sampling. In practice, the reverse process variance $\sigma_t^2$ is often fixed to either $\beta_t$ or $\tilde{\beta}_t$, as proposed in the original DDPM formulation [47].

The denoising step is repeated from $t = T$ down to $t = 0$ until the final denoised sample $x_0$, representing a synthetic data sample, is obtained.

### MRI-to-CT Transformer-based Improved Denoising Diffusion Probabilistic Model (MC-IDDPM)

The MC-IDDPM framework, introduced in [22], extends the standard DDPM formulation to the task of synthesizing CT images from a patient's MRI. While the conventional DDPM [47] is unconditional, predicts only the mean of the reverse Gaussian distribution, and often relies on convolutional U-Net architectures, MC-IDDPM is conditional on the paired MRI, predicts both the mean and variance of the reverse distribution for more accurate denoising, and employs a shifted-window Vision Transformer-based V-Net (Swin-VNet) to capture long-range dependencies and global context more effectively. These modifications allow MC-IDDPM to substantially reduce the number of inference steps while improving synthesis accuracy, outperforming multiple state-of-the-art GAN- and DDPM-based methods across both brain and prostate datasets. including MC-GAN [49], MC-CGAN [49], 2D-IDDPM [50], 3D-DDIM [51], and 3D-DDPM [52]. MC-IDDPM demonstrated improvements in most evaluation metrics, while requiring substantially fewer inference steps, making it more accurate and efficient for MRI-to-CT translation.

The diffusion process consists of three steps as follows:

#### Forward Diffusion

The forward process is identical to the DDPM formulation in Eq. 5.1, with a linear variance schedule $\beta_t = 6 \times 10^{-6} \cdot t$ and a total of 1000 timesteps were used for training, with 50 sampled during inference.

#### Reverse Diffusion

Unlike the original DDPM, which fixes $\Sigma_\theta$, MC-IDDPM jointly learns both the mean $\mu_\theta$ and variance $\Sigma_\theta$ of the reverse distribution, conditioned on the MRI input $Z$.

As in DDPM, the network predicts the $\epsilon_\theta$, but uses an MAE objective rather than MSE.

The predicted mean can be recovered using the predicted noise:

$$\mu_\theta(x_t, t|Z) = \frac{1}{\sqrt{1-\beta_t}} \left( x_t - \frac{\beta_t}{\sqrt{1 - \Sigma_{i=1}^{t} 1 - \bar{\beta}_i}} \epsilon_\theta(x_t, t|Z) \right) \tag{3.14}$$

To estimate the variance, the network learns an interpolation factor $k_\theta(x_t, t|Z)$, and defines:

$$\Sigma_\theta(x_t, t|Z) = \exp(k_t(x_t, t|Z)\log\beta_t + (1 - k_t(x_t, t|Z)) \times \log(\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t) \tag{3.15}$$

The coefficient $k_\theta$ is learned by minimizing a variational lower bound loss:

$$\arg\min_{k_\theta} L_{\text{var}} = \arg\min_{k_\theta} \mathcal{L}_{\text{VLB}}(\Sigma_t, \Sigma_\theta(x_t, t|Z)) \tag{3.16}$$

Finally, the total loss function combines both mean and variance objectives:

$$L = L_{\text{mean}} + \gamma L_{\text{var}} \tag{3.17}$$

where $\gamma$ is a weighting factor set based on the ratio between inference and training steps (e.g., $\gamma = 0.0125$). $L_{\text{mean}}$ is the mean absolute error (MAE) loss used for noise prediction, and $L_{\text{var}}$ is the variational lower bound loss used for variance estimation.

#### Generating sCT

After training, the estimated $\mu_\theta$ and $\Sigma_\theta$ are used in Equation 3.17 to iteratively denoise $x_t$ and recover the clean image $x_0$. During inference, 50 timesteps are uniformly sampled from $t = 1$ to $t = 1000$, forming a set $s \in \{s_1, s_2, \ldots, s_{50}\}$. Starting from Gaussian noise at $x_{s_{50}} \sim \mathcal{N}(0, I)$, the model applies the reverse process through the selected timesteps to reconstruct the final synthetic CT image $x_0$ aligned with the input MRI.

Due to the stochasticity of $\epsilon$ at each step, the generation is repeated multiple times per patient and averaged to produce a stable sCT output, following a Monte Carlo-style sampling strategy.

Although this approach improves synthesis efficiency by reducing the number of generation timesteps, it remains computationally expensive compared to GANs.

### 3.2.3 Image Similarity Metrics for Synthesis Evaluation

To evaluate the quality of the generated images, several commonly used metrics are summarized below.

#### Mean Absolute Error (MAE)

MAE measures the average absolute voxel-wise difference between the sCT and the real CT. It provides a direct measure of voxel-level intensity accuracy [53].

$$\mathrm{MAE}(\mathrm{CT}, \mathrm{sCT}) = \frac{1}{N} \sum_{i=1}^{N} |\mathrm{CT}_i - \mathrm{sCT}_i| \tag{3.18}$$

Here, $N$ is the total number of voxels, and $\mathrm{CT}_i$, $\mathrm{sCT}_i$ are the voxel intensities at index $i$ [53].

### Peak Signal-to-Noise Ratio (PSNR)

PSNR calculates the peak signal value to the mean squared error [53]. higher values indicate better intensity agreement.

$$\mathrm{PSNR}(\mathrm{CT}, \mathrm{sCT}) = 10 \cdot \log_{10} \left( \frac{Q^2}{\frac{1}{N} \sum_{i=1}^{N} (\mathrm{CT}_i - \mathrm{sCT}_i)^2} \right) \tag{3.19}$$

Where $Q$ denotes the maximum possible voxel intensity value in the CT image, and $N$ is the total number of voxels [53]. Higher PSNR indicates lower pixel-wise error.

### Structural Similarity Index Measure (SSIM)

SSIM evaluates the structural similarity between CT and sCT and is defined as follows. [53]:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3.20}$$

Where $\mu_x, \mu_y$ are local means, $\sigma_x^2, \sigma_y^2$ are local variances, $\sigma_{xy}$ is the local covariance, and constants $c_1 = (0.01 \cdot L)^2$, $c_2 = (0.03 \cdot L)^2$, with $L$ as the intensity dynamic range [53].

The final SSIM score is:

$$\mathrm{SSIM}(\mathrm{CT}, \mathrm{sCT}) = \frac{1}{N} \sum_{i=1}^{N} \mathrm{SSIM}(\mathrm{CT}_i, \mathrm{sCT}_i) \tag{3.21}$$

### Pearson Correlation Coefficient (PCC)

PCC measures the linear correlation between the voxel intensities of the sCT and the real CT. It evaluates how well the intensity values of sCT match the trends in the real CT. A value of 1 indicates perfect positive correlation, while 0 indicates no correlation.

$$\mathrm{PCC}(\mathrm{CT}, \mathrm{sCT}) = \frac{\sum_{i=1}^{N} (\mathrm{CT}_i - \bar{\mathrm{CT}})(\mathrm{sCT}_i - \bar{\mathrm{sCT}})}{\sqrt{\sum_{i=1}^{N} (\mathrm{CT}_i - \bar{\mathrm{CT}})^2} \sqrt{\sum_{i=1}^{N} (\mathrm{sCT}_i - \bar{\mathrm{sCT}})^2}} \tag{3.22}$$

Here, $\bar{\text{CT}}$ and $\bar{\text{sCT}}$ represent the mean intensity values of the CT and sCT, respectively.

**Dice Similarity Coefficient (DSC)**

The DSC evaluates the spatial overlap between two binary segmentations between ground-truth structures from the real CT and predicted segmentations from the sCT. It is particularly useful in assessing anatomical accuracy in the predicted segmentation.

$$\text{DCS} = \frac{2TP}{2TP + FP + FN} \tag{3.23}$$

where $TP$ denotes the number of true positive voxels (correctly identified as belonging to the structure of interest), $FP$ represents false positive voxels, and $FN$ represents false negative voxels. A higher DSC indicates better spatial overlap and segmentation accuracy. A score of 1 indicates perfect agreement between the two segmentations, whereas a score of 0 indicates no overlap.

In this study, we employed TotalSegmentator [54] to extract anatomical regions from our synthesized CT images for DSC evaluation. It is a DL toolkit based on nnU-Net [55], trained on 1204 clinical CT scans for multi-structure segmentation. It covers 104 anatomical structures across CT and MRI and achieved an average DSC of 0.943 on its CT test set. It provides detailed segmentation for clinically relevant head and neck structures, including the left and right parotid glands and the spinal cord, which are crucial for radiotherapy planning. Its broad anatomical coverage and generalization performance made it a tool for inference-time segmentation.

### 3.2.4 Dose evaluation metrics

The dose plan from real CT and sCT images enables comparison of the clinical evaluation of the sCT against that of the real CT. $\text{Dose}_{MAE}$ and Dose Volume Histogram (DVH) difference are metrics used to evaluate the predicted dose from sCTs. $\text{Dose}_{MAE}$ evaluates the average absolute difference between the predicted dose obtained from the real CT and from the sCT. DVH difference is the average absolute difference between two predicted DVHs across all structures (PTVs and OARs), showing how much the dose prediction changes when using sCT instead of the real CT.

In this study, Swin UNETR++ [56], a transformer-based DL model for dense 3D radiation dose prediction, is employed to predict the dose for both real and sCTs. It was trained on the publicly available OpenKBP dataset [57], which contains CT scans and dose distributions

from 340 HNC cancer patients. Each case includes three (PTVs)—PTV70, PTV63, and PTV56—representing high-, intermediate-, and low-risk tumor regions prescribed with 70, 63, and 56 Gy, respectively. Additionally, seven OARs, including the brainstem, spinal cord, parotid glands, larynx, and mandible, are delineated to minimize radiation exposure to healthy tissue during planning.

## 3.3 Medical Image Registration

In medical image analysis, image registration aligns two or more images within a common spatial coordinate system to match corresponding anatomical structures. This alignment enables the assessment of changes in shape, size, or position over time or across individuals, supporting disease tracking and treatment planning. The primary goal is to determine the optimal transformation that maps the moving image onto the fixed image, ensuring reliable results for further analysis [58].

### 3.3.1 Rigid Registration

Rigid registration aligns images by applying only rotations and translations, without altering their scale or shape. It is most effective when differences between scans arise purely from patient movement or changes in positioning, such as head repositioning between imaging sessions. Because this method preserves the original geometry and size of anatomical structures, it ensures accurate spatial correspondence when no deformation has occurred. However, it cannot account for anatomical shape variations or local structural changes between subjects or over time [58].

### 3.3.2 Affine Registration

Affine registration expands on rigid transformation by allowing not only translation and rotation, but also scaling and shearing. This is often used as an initial step for deformable registration to resolve global misalignments [58].

### 3.3.3 Deformable (Non-Rigid) Registration

Deformable transformation is the most flexible, allowing local, point-wise displacements through a dense vector field. It accounts for anatomical variability and deformation due to changes (e.g., tumor growth, organ compression). Such transformations are particularly valuable in applications that demand high local accuracy, for instance, adaptive radiother-

apy, where rigid or affine methods fail to capture subtle, non-uniform tissue changes. One example is the alignment of pre- and post-treatment scans in oncology, where it is essential to accurately represent small-scale tissue shifts or shrinkage [58]. Some state-of-the-art deformable medical image registration methods are described in the following sections.

### (1) Classical Deformable Registration

Classical deformable registration methods formulate the problem as the minimization of an energy functional that balances image similarity with deformation regularity. Given a fixed image $F$ and a moving image $M$, the goal is to estimate a deformation field $\phi$ such that the warped moving image $M(\phi)$ is well aligned with $F$. This can be written as:

$$E(\phi) = D\Big(F, M(\phi)\Big) + \alpha\, R(\phi), \tag{3.24}$$

where $D(\cdot,\cdot)$ is a dissimilarity or similarity term, $R(\phi)$ is a regularization term promoting smooth and plausible deformations, and $\alpha > 0$ controls the trade-off between alignment accuracy and smoothness.

In many medical image registration frameworks, non-rigid deformation fields are modeled using cubic B-spline Free-Form Deformations (FFD). In this parameterization, implemented in tools such as NiftyReg [59], the transformation $\phi$ is defined by a grid of control points, and the deformation at each voxel is obtained by interpolating these points with cubic B-spline basis functions. This yields a smooth and efficient non-rigid deformation field, well suited for large 3D registrations such as CT–CBCT.

Other classical approaches include Demons-based methods [60], which iteratively update dense displacement fields using forces derived from image gradients, and diffeomorphic frameworks such as LDDMM [61], which explicitly enforce invertibility and topology preservation. Regularization typically takes the form of diffusion or bending-energy penalties, e.g., $\|\nabla\phi\|^2$ or $\|\nabla^2\phi\|^2$, which suppress unrealistic local distortions. Although effective and theoretically grounded, these methods require iterative optimization for each image pair, resulting in high computational cost in large-scale 3D datasets.

### Similarity Measures and Cost Functions

While transformation models determine how the moving image can be aligned with the fixed one, the accuracy of registration depends on how well this alignment is measured. Similarity metrics provide this measure by serving as the objective function that guides the optimization

toward better correspondence. Depending on the modality and registration task, different metrics are used to quantify how closely the transformed moving image matches the fixed image [58].

Several similarity metrics can quantify how well the transformed moving image matches the fixed one. A simple option is the Sum of Squared Differences (SSD), which measures voxel-wise intensity differences, but it is highly sensitive to noise and intensity variations, making it less reliable clinically [58]. Normalized Cross-Correlation (NCC) improves robustness by comparing the relative patterns of intensity variation rather than absolute values. By subtracting the mean and normalizing by the variance, NCC is more robust to linear intensity shifts and is commonly used in unimodal registration tasks where structural correspondence is preserved despite contrast changes [58]. For multimodal registration, Mutual Information (MI) is commonly used because it measures statistical dependence between intensity distributions rather than direct intensity similarity, enabling alignment across modalities such as CT and MRI. However, MI requires good initialization [58].

**(2) Deep Learning-based Deformable Registration**

Deep learning has reshaped deformable registration by replacing per-pair iterative optimization with a learned model that directly predicts deformation fields from a fixed–moving image pair. Early approaches were fully supervised and trained on synthetic or simulated deformation fields [62,63]. While these models performed well in controlled settings, their applicability to real clinical data was limited by the lack of reliable ground-truth deformations.

A major step forward came with end-to-end unsupervised registration models, which eliminated the need for ground-truth deformation fields. Methods like the DIRNet [64] showed that a network can learn to predict smooth deformation fields directly from image pairs by optimizing image-similarity and regularization losses. These frameworks also integrated differentiable spatial transformers, allowing the warping operation to be learned jointly with the deformation prediction.

Subsequent research introduced more powerful architectures. CNN-based U-Net variants became widely adopted for their ability to capture both global anatomical context and fine-grained spatial details [64]. Semi-supervised extensions incorporated sparse supervision from anatomical labels or landmarks to improve anatomical plausibility, especially when training data were limited [65]. Other work introduced modality-robust similarity measures, attention mechanisms, or multi-resolution strategies to better handle multimodal challenges such as CT–CBCT alignment [66].

More recently, transformer-based models have been proposed to address the limitations of CNNs in modeling long-range spatial relationships. Early transformer registration frameworks such as TransMorph [67] demonstrated that attention-based global context can improve alignment robustness, especially in anatomically complex or low-contrast regions.

The deep learning registration methods used in this study, VoxelMorph, PC-Reg-RT, and XMorpher, build upon these foundations. They represent three influential directions in deep learning registration: unsupervised UNet-based registration, semi-supervised perception–correspondence strategies, and transformer-based cross-attention architectures. These models are described in detail in the following sections.

### VoxelMorph

VoxelMorph [8] is a learning-based framework that leverages convolutional neural networks (CNNs) to perform fast and accurate deformable image registration.

VoxelMorph introduces an unsupervised learning paradigm for efficient 3D deformable medical image registration. Unlike traditional approaches that independently optimize a transformation for each image pair, VoxelMorph learns a global parametric registration function $g_\theta$, implemented as a CNN. This function takes a pair of affine-aligned volumes, the moving image $M$ and the fixed image $F$, as a 2-channel input and outputs a dense deformation vector field (DVF) $\phi$, representing voxel-wise displacement vectors that map each voxel in $F$ to its corresponding location in $M$.

The CNN architecture in VoxelMorph follows a U-Net-like [6], effectively capturing both global context and fine-grained spatial features. The output DVF is optimized during training using a spatial transformer network (STN), which computes warped images $M(\phi)$ and backpropagates errors through the warping operation. The full pipeline of the VoxelMorph method is illustrated in Figure 3.5. The figure shows how the moving and fixed 3D images are input to the model to estimate the deformation field $\phi$, which is then used by a STN to warp the moving image. The resulting aligned image $M(\phi)$ is compared to the fixed image $F$ through a loss function that combines image similarity and smoothness of the deformation. The deformation field $\phi$ can also be used to warp the segmentations and dose plan corresponding to the moving image.

The loss function used for training is defined as:

$$L(F, M, \phi) = L_{\text{sim}}(F, M(\phi)) + \lambda L_{\text{smooth}}(\phi) \tag{3.25}$$

Here, $\lambda$ is a regularization weight that balances the similarity term $L_{\text{sim}}$ and the smoothness term $L_{\text{smooth}}$. The similarity term encourages alignment between the fixed image $F$ and the warped moving image $M(\phi)$. VoxelMorph uses the negative local cross-correlation (CC) as $L_{\text{sim}}$, which is robust to intensity variations and evaluates local structural similarity across small neighborhoods.

The smoothness term $L_{\text{smooth}}$ promotes spatial coherence in the deformation field by penalizing abrupt changes or non-realistic distortions. It is defined as:

$$L_{\text{smooth}}(\phi) = \sum_{p \in \Omega} \|\nabla \phi(p)\|^2 \tag{3.26}$$

where $\nabla \phi(p)$ denotes the spatial gradients of the deformation field at voxel $p$, and $\Omega$ represents the spatial domain of the image. This term is implemented as an $L_2$-norm over the voxel-wise gradient magnitudes, enforcing a smooth and plausible transformation across the entire image.

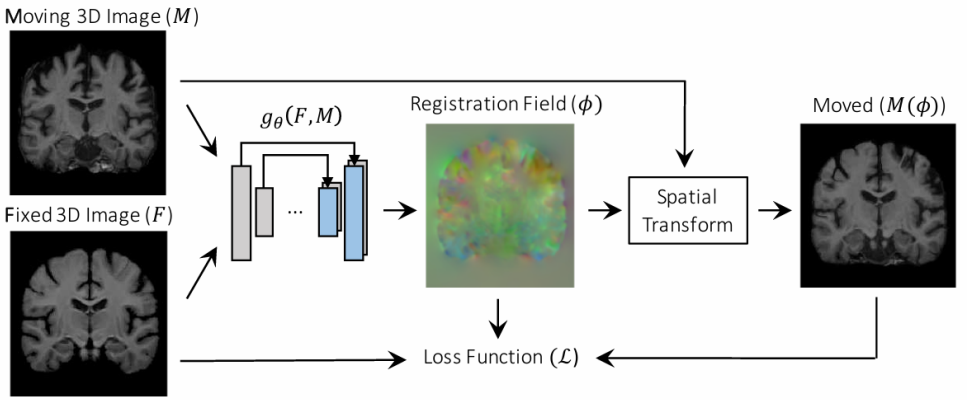


Figure 3.5 Overview of the VoxelMorph pipeline. The network learns parameters for a function $g$ that registers a moving volume $M$ to a fixed volume $F$. During training, the deformation field $\phi$ is used to warp $M$ via a spatial transformer. The loss function compares $M(\phi)$ with $F$ and promotes smoothness in $\phi$. The image is taken from [8].

### Perception-Correspondence Registration with Reverse Teaching (PC-Reg-RT)

While VoxelMorph offers an efficient unsupervised approach for deformable registration, it lacks the ability to perceive specific anatomical regions. This leads to two key issues: (1) misalignment of low-contrast or blurry anatomical structures, and (2) distortion in task-irrelevant regions [9]. Although label-constrained models attempt to address this by embedding per-

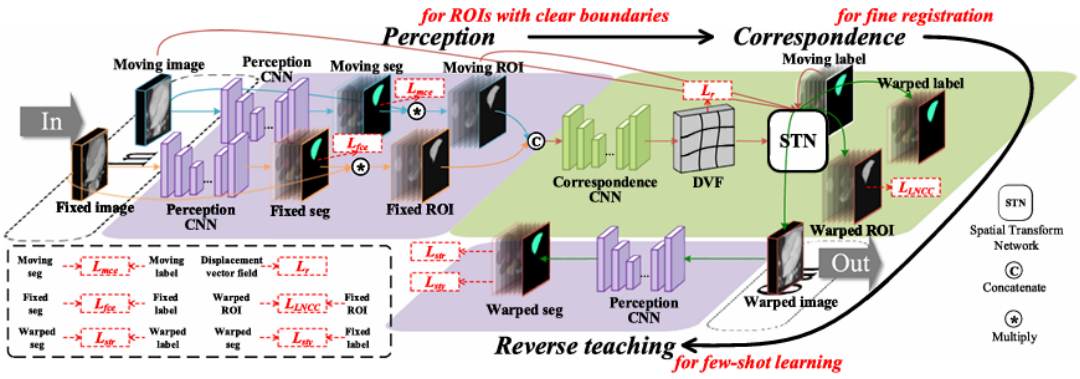


Figure 3.6 Overview of the PC-Reg-RT framework. The registration task is decomposed into two stages: perception and correspondence. The perception CNN extracts anatomical ROIs with clear boundaries, while the correspondence CNN performs accurate texture-preserving registration. Reverse Teaching enriches the perception CNN with structural and style knowledge from unlabeled images, enabling few-shot learning. The image is taken from [9].

ception through segmentation labels, they are limited by high annotation costs [9].

To overcome these challenges, PC-Reg-RT [9] leverages a semi-supervised registration framework. The core idea is to decouple the registration task into two independently optimized tasks: (1) a Perception CNN, which segments regions of interest (ROIs) to identify anatomical boundaries, and (2) a Correspondence CNN, which performs deformable registration specifically on those segmented ROIs. This separation avoids interference between semantic segmentation and spatial alignment, thereby reducing deformation-related distortions and improving accuracy, particularly in few-shot scenarios with limited labels [9].

Training in PC-Reg-RT follows a two-stage strategy. In the first stage, the Correspondence CNN is trained with a small number of manually labeled moving images to produce initial ROI segmentations for both fixed and moving volumes. The Perception CNN then computes a deformation field that aligns only the ROI regions rather than the entire image, allowing for more anatomically focused and texture-preserving registration. Although only the moving images are partially labeled, the segmentation model is supervised effectively using these labels, benefiting from the fact that both labeled and unlabeled inputs belong to the same imaging modality. As a result, the segmentation model generalizes well to the unlabeled fixed images during training, despite being directly optimized only on the labeled moving ones.

In the second stage, PC-Reg-RT introduces its key innovation: the Reverse Teaching mechanism. Here, moving images and their segmentations are warped to match a diverse set of fixed, unlabeled images using the Perception CNN. This generates augmented training pairs where the original segmentations are now exposed to a wider range of anatomical structures

and imaging styles. These new pairs are then used to retrain the Correspondence CNN, enriching its ability to generalize beyond the few annotated samples. As the Perception CNN improves, it provides more accurate and robust ROIs for the Correspondence CNN, leading to a positive feedback loop that enhances overall registration performance.

By combining ROI-specific registration, independent optimization of segmentation and deformation, and an effective few-shot learning strategy, PC-Reg-RT achieves accurate, low-distortion registration with minimal annotation effort, making it well-suited for clinical applications with limited labeled data [9]. An illustration of the overall PC-Reg-RT framework, including the decoupled CNN modules and Reverse Teaching strategy, is shown in Fig. 3.6.

The network is optimized using a local normalized cross-correlation (NCC) loss with a uniform window size of 9 between the warped CT and the fixed CBCT, combined with a regularization term to promote smooth deformations, and supplemented by a cross-entropy loss between the predicted and reference labels when available.

### XMorpher

XMorpher [10] is a full-transformer-based architecture. Unlike models such as VoxelMorph [8], which employ a single U-Net-like [6] backbone processing concatenated inputs, XMorpher uses a dual-path strategy (Figure 3.7) where the fixed image $f$ and the moving image $m$ are processed by two parallel U-shaped networks. These networks exchange information at multiple levels using a novel Cross Attention Transformer (CAT) block, enabling progressive and symmetric feature fusion. This fusion-first approach, used by VoxelMorph, often causes feature distortion within mixed regions [10].

Each CAT block computes attention between base and searching windows derived from local regions of $f$ and $m$, facilitating efficient learning of spatially constrained voxel-wise correspondences. By employing multi-size window partitioning, the network focuses on local displacements while improving computational efficiency and avoiding incorrect long-range matches.

As illustrated in Figure 3.7, the extracted features from both branches are progressively merged using attention-guided fusion. The training is guided by a similarity loss $\mathcal{L}_{\text{sim}}$, based on normalized cross-correlation (NCC) which is computed locally (within a uniform window size of 9), i.e., within a sliding window rather than over the entire image, and a smoothness regularizer $\mathcal{L}_{\text{smooth}}$, same to the one used in VoxelMorph.

XMorpher is applied in two training settings:

1. **Unsupervised (VM-XMorpher):** Built on the VoxelMorph [8] paradigm, this vari-

ant uses only image similarity and regularization losses to learn $\phi$, without requiring segmentation labels.

2. **Semi-supervised (PC-XMorpher):** This variant builds on the PC-Reg-RT [9] framework by incorporating the XMorpher transformer as the registration backbone while retaining the PC-Reg-RT architecture for the segmentation component.

In the unsupervised setting, compared to VoxelMorph and TransMorph [67], VM-XMorpher achieves a higher Dice Similarity Coefficient (DSC), demonstrating its superior ability to model fine-grained, anatomically consistent deformations. In the semi-supervised setting, PC-XMorpher also outperforms PC-Reg-RT, achieving more accurate alignment of anatomical structures by incorporating the cross-attention mechanism. These improvements highlight the effectiveness of XMorpher's dual-path cross-attention architecture in learning robust correspondences between moving and fixed images for deformable medical image registration [10].

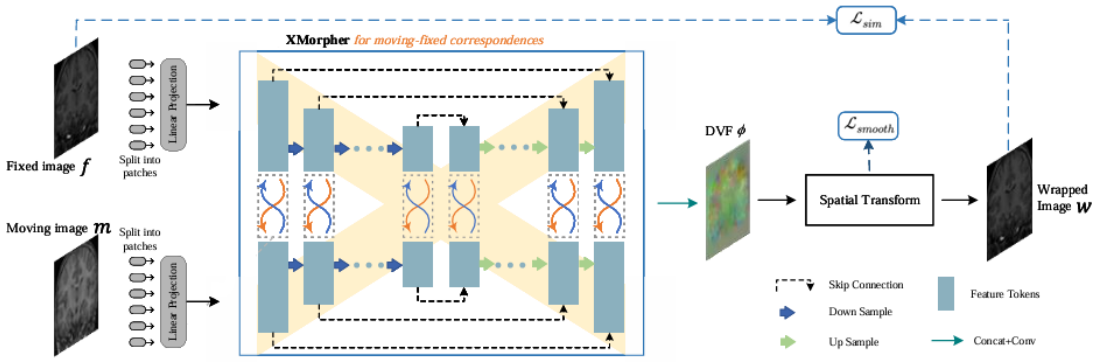


Figure 3.7 Overview of the XMorpher architecture. Dual U-Net-like transformer branches extract and fuse features from fixed and moving images using window-based cross-attention mechanisms, producing a deformation field $\phi$ optimized via similarity and smoothness losses. The image is taken from [10].

### Challenges and Limitations of Deep Learning–based Registration Methods

One major limitation of state-of-the-art methods like PC-XMorpher and PC-Reg-RT is that their evaluations have been restricted to unimodal registration tasks, PC-XMorpher on cardiac CT-to-CT and PC-Reg-RT on both cardiac CT-to-CT and brain MRI-to-MRI.A key challenge arises when the imaging modalities differ, as in our case, where the moving image is a CT scan with available segmentations, while the fixed image is an unlabeled CBCT. The

perception network, which is similar in both models, struggles to segment the lower-quality CBCTs in which structural boundaries are less visible compared to the higher-quality CTs.

### 3.3.4 Registration Evaluation metrics

#### Normalized Cross Correlation (NCC)

NCC compares the similarity of voxel intensities between the fixed and moved images after subtracting their means, offering robustness to global intensity shifts. It also serves as a similarity cost function for evaluating how well two images are aligned in registration [58].

$$\mathrm{NCC}(f, m) = \frac{\sum_{x \in D} (f(x) - \mu_f)(m(x) - \mu_m)}{\|f - \mu_f\|_2 \cdot \|m - \mu_m\|_2} \tag{3.27}$$

Where x represents the voxel location within the image domain D, $f$ is the fixed image, $m$ is the moved (transformed) image, $\mu_f$, $\mu_m$ are their mean intensities of images, $\|\cdot\|_2$ denotes the $\ell_2$-norm (Euclidean distance) over the image domain [58].

#### Target Registration Error (TRE)

TRE quantifies how well a transformation aligns corresponding landmarks between two images. It is commonly applied when manual landmark pairs are available [68]. The TRE is defined as:

$$\mathrm{TRE} = \frac{1}{N} \sum_{i=1}^{N} \left\| T(l_m^i) - l_f^i \right\|_k \tag{3.28}$$

where $T$ is the estimated transformation that maps points from the moving image to the fixed image. $l_m^i$ and $l_f^i$ are the $i$-th corresponding landmarks in the moving and fixed images, respectively. $\|\cdot\|_k$ denotes the distance metric, which can be either the $\ell_1$-norm (Manhattan distance) or the $\ell_2$-norm [68].

#### Deformation Regularity and Jacobian Determinant

In addition to image similarity and landmark accuracy, deformation regularity provides important information about the plausibility of the estimated transformation. Smoothness is often quantified using gradient-based penalties such as the diffusion term $\|\nabla\phi\|^2$ or bending-energy measures $\|\nabla^2\phi\|^2$, which capture abrupt local variations in the displacement field [69].

Topology preservation is commonly assessed using the Jacobian determinant $J_\phi(x)$, which characterizes local volume change. Voxels with $J_\phi(x) \leq 0$ indicate foldings or non-invertible mappings, and reporting the percentage of such voxels has become standard practice in recent registration quality-assessment studies [70].

## 3.4 Summary

In this chapter, we established the foundational concepts of deep learning methods applied to medical image synthesis and deformable registration, outlining the underlying principles, representative approaches, and their respective limitations, such as GAN-based approaches for CT synthesis, which face difficulties in preserving fine anatomical structures due to the reliance on L1 or L2 losses within the adversarial framework, a key limitation for downstream tasks such as dose prediction, segmentation, and registration essential in ART, and semi-supervised registration models, which leave multimodal alignment largely unexplored. Moreover, aligning the dose plan with daily CBCTs is a critical step that facilitates ART, and this study further aims to leverage and adapt state-of-the-art registration networks using cross-attention to improve registration accuracy, thereby mitigating global alignment limitations and enabling fine alignment, specifically in the GTV and OARs. This foundation provides the necessary context for introducing the proposed methodology, which will be presented in the following chapter. A unified MRI-based ART approach combines MRI-to-CT synthesis with deformable CT-to-CBCT registration, addressing structural inconsistencies in synthesis and the multimodal CT-CBCT registration. This enables better exploitation of MRI information within ART. By integrating MRI into CBCT workflows through our approach, current limitations can be mitigate, improving anatomical fidelity and supporting more accurate and reliable adaptive treatment.

# CHAPTER 4 PROBLEM STATEMENT, HYPOTHESIS, RESEARCH OBJECTIVES AND METHODOLOGY

The previous chapters 2 and 3 outlined the clinical challenges in HNC radiotherapy and reviewed deep learning solutions for treatment adaptation. This study focuses on two components for deformable registration of MRI to daily CBCTs: (1) synthesizing CT images from MRI, and (2) achieving alignment between the synthetic CT (sCT)/MRI and daily CBCT scans to track anatomical changes throughout the treatment, and can support MRI-based adaptive radiotherapy (ART).

## 4.1 Problem Statement

An MRI-based framework that enables ART is a promising approach that has the potential to eliminate separate CT scans and enable adaptation to daily anatomical changes. However, high-fidelity sCT generation from MRI is still required for accurate dose planning, segmentation, and registration. Deformable registration between sCT and daily CBCT scans is challenged by cross-modality discrepancies, and in semi-supervised methods, it is further hindered by the lack of annotations in low-quality, artifact-prone CBCTs. While labeled CT scans provide reliable supervision, the absence of clinically usable ground truth in CBCT limits the feasibility of direct semi-supervised registration.

## 4.2 Hypothesis

A unified framework that first synthesizes a patient's CT from MRI and then registers the sCT to each daily CBCT has the potential to enable MRI-based ART. In the synthesis stage, incorporating structural constraints within a generative model can improve the quality of the sCT, while diffusion-based synthesis can provide higher-fidelity inputs for the registration stage. Transformer-based deformable CT-to-CBCT registration with cross-attention can then enable more accurate image alignment.

## 4.3 Objective

The objective of this project is to develop and validate a deep learning–based pipeline for MRI-to-CBCT deformable registration in HNC, enabling MRI-driven adaptive radiotherapy workflows. The Sub-objectives are: 1. Develop an adversarial synthesis network enhanced

with a feature consistency loss to generate sCTs with improved structural fidelity, and leverage a diffusion-based model at test time, to generate sCTs with higher image quality for the subsequent registration task. 2. Adapt and evaluate registration networks to achieve accurate alignment between sCTs/CT and daily CBCTs, ensuring reliable anatomical correspondence. 3. Assess the clinical relevance of the synthesis pipeline by evaluating dose metrics.

## 4.4 Data Acquisition and Pre-processing

This study employed two datasets: The public SynthRad2023 brain dataset [1] and an institutional head and neck cancer (HNC) cohort from the Centre Hospitalier de l'Université de Montréal.

### 4.4.1 SynthRad2023 brain dataset

This dataset includes 180 paired MRI/CT public cases from patients treated with radiotherapy at three centers. We performed 5-fold cross-validation, ensuring that each fold contained equal proportions of data from all three centers, with 80% used for training and 20% for validation. Images were cropped to a uniform size of 192 × 192 × 164 and normalized to [–1, 1] for synthesis.

### 4.4.2 HNC dataset

The HNC cohort consists of 338 patients treated with curative-intent radiotherapy. Each patient had a pCT, associated segmentations and dose plan, one MRI, and multiple CBCT scans. Of these, 282 patients were used for 5-fold cross-validation, and 56 were reserved for testing. MRI/CT images were resampled to an isotropic voxel spacing of 1 mm, while CT/CBCT pairs were resampled to an isotropic voxel spacing of 2 mm using nearest neighbour interpolation.

Rigid and affine registrations were performed using Elastix v5.2.0 [71] with four resolution levels. Each pCT, along with its contours and dose plan, was first affinely aligned to the initial CBCT ($CBCT_0$) using normalized cross-correlation (NCC), followed by MRI to aligned pCT registration with normalized mutual information. The remaining CBCTs were rigidly aligned to $CBCT_0$. Prior to registration, CTs were cropped around the tumor volume. Intensity normalization was applied within the head and neck regions based on their mean and standard deviation. MRI/CT inputs for synthesis were cropped to 256 × 256 × 79 voxels, and CT/CBCT inputs for registration were resampled to 128 × 128 × 128 voxels. For synthesis, MRI/CT intensities were normalized to [–1, 1], and for registration, CT/CBCT to [0,

1]. Figure 4.1 summarizes the pre-processing registration flow in HNC dataset.

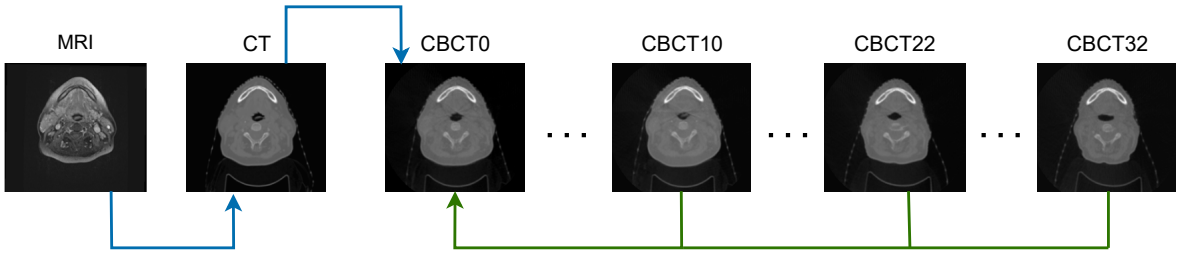


Figure 4.1 HNC pre-processing. The figure shows the affine and rigid registration steps used in preprocessing, indicated by blue and green arrows, respectively. First, the CT is affine-registered to $CBCT_0$, then the MRI is affine-registered to the aligned CT, and finally, all other CBCTs are rigidly registered to $CBCT_0$.

## 4.5 Methodology Overview

The methodology developed in this work consists of a two-stage pipeline designed to enable MRI-based ART through deformable registration of MRI-derived sCTs to daily CBCT scans. The first stage addresses MRI-to-CT synthesis, where sCTs are generated to provide CT-equivalent information for dose calculation and segmentation. The second stage focuses on deformable registration, aligning the CT with daily CBCTs. Together, these stages establish a framework for multimodal alignment and support the integration of MRI into adaptive workflows for HNC. At each stage, several methods were used or developed, and the best-performing method from each stage was selected for the final pipeline. Several methods were adopted or developed at each stage, and the best-performing method from each stage was utilized in the final pipeline.

### 4.5.1 Stage 1: MRI-to-CT synthesis

We began the synthesis stage with a conditional generative adversarial network (CGAN) and subsequently enhanced it into a feature-consistency CGAN (FCGAN) by incorporating a feature-consistency loss term into the baseline CGAN generator objective, as detailed in the following section. We also adapted the diffusion-based method MC-IDDPM within our synthesis pipeline, and for the end-to-end MRI-to-CBCT pipeline, we selected the best-performing synthesis method.

### Gan-based

This stage of the pipeline aims to generate anatomically accurate sCT images from patient MRIs using the enhanced FCGAN model, which incorporates a feature-consistency loss to better preserve structural details.

- Network Design and Training Strategy

  The baseline CGAN is inspired by the Pix2Pix framework [38], with a U-Net style generator [6] and a PatchGAN discriminator. The generator encodes MRI information through successive convolutional layers, compressing spatial features before reconstructing the CT image using mirrored upsampling layers. Skip connections are incorporated to preserve fine-grained spatial details. The discriminator operates at the patch level, distinguishing local regions of real CT images from synthetic ones.

  In the baseline CGAN, the generator is optimized using a combination of binary cross-entropy loss from the adversarial training and an $\ell_1$ reconstruction loss between the predicted CT and the corresponding ground-truth CT [38]. FCGAN extends this approach by introducing an additional feature consistency term into the generator loss, which enforces alignment between high-level features extracted from the input MRI and those obtained from real CT images. Specifically, this loss is defined as an $\ell_1$ distance between the feature map from the final layer of the generator's encoder, to which a 2D convolution layer is applied to adjust the output dimensionality to match the discriminator's feature map output and the feature representation extracted by the discriminator from the real CT. This additional constraint complements the adversarial and voxel-wise losses, guiding the generator to capture the structural characteristics of real CT anatomy before sCT generation.

  For both the baseline CGAN and FCGAN, the discriminator is trained with an $\ell_1$ loss, encouraging it to assign outputs close to zero for sCT–MRI pairs and close to one for real CT–MRI pairs.

### Diffusion-based

For CT synthesis, we also used the diffusion model MC-IDDPM described in Section 3.2.2 and followed its approach, with a larger input patch size compared to the original setting, as detailed in Section 4.5.1, to achieve improved synthesis performance.

### Synthesis Evaluation

For evaluating the synthesized CTs produced by the CGAN, FCGAN, and MC-IDDPM models, we employed standard image similarity metrics, including the mean absolute error (MAE), peak signal-to-noise ratio (PSNR), pearson correlation coefficient (PCC), and structural similarity index (SSIM). The initial comparison on SynthRad2023 employed 5-fold cross-validation, where CGAN and FCGAN were evaluated against the diffusion-based MC-IDDPM, which was ultimately adopted as the final synthesis model. To further validate model generalizability, all synthesis methods were subsequently evaluated on the HNC dataset, also using 5-fold cross-validation. On this cohort, the same image similarity metrics were computed, and anatomical fidelity was additionally quantified using the Dice similarity coefficient (DSC) between ground-truth CT segmentations and those derived from TotalSegmentator [54] applied to the sCT volumes.

To clinically evaluate the sCTs on the HNC dataset across the three synthesis methods, a dosimetric assessment was performed using the MAE of the dose distribution and the dose–volume histogram (DVH) differences between doses calculated on real CTs and those on sCTs. Dose calculations for both real and generated CTs followed the approach proposed by [56] (described in Section 3.2.4), with dose MAE computed over the entire head-and-neck region. DVH differences were quantified in two ways: (1) an overall DVH score obtained by aggregating errors across all planning target volumes and organs at risk (OARs), and (2) per-structure DVH metrics computed individually for each target and OAR.

### Experimental details

The CGAN and FCGAN were trained for 100–240 epochs using the Adam optimizer, with the number of epochs determined by the best image similarity metrics (MAE, PCC, PSNR, and SSIM). A learning rate of $1 \times 10^{-5}$ was selected after evaluating candidate values in the range $1 \times 10^{-5}$ to $1 \times 10^{-4}$. A batch size of 8 was adopted after testing values of 1, 8, and 16. Data augmentation included random flips and random affine transformations with a probability of 50%, implemented using TorchIO [72].

For MC-IDDPM, we followed the original implementation with the modification of increasing the patch size to $192 \times 192 \times 4$ for SynthRad2023 and $256 \times 256 \times 4$ for HNC, as the smaller sub-volumes ($64 \times 64 \times 4$) used in the original study yielded suboptimal results.

### 4.5.2 Stage 2: CT-to-CBCT deformable Registration

For multimodal deformable alignment between CTs/sCTs and CBCTs, we use the XMorpher family [10], adapting both its unsupervised and semi-supervised variants to head-and-neck CT–CBCT.

#### Deformable registration with XMorpher

VM-XMorpher serves as our unsupervised baseline for multimodal CT-to-CBCT registration. The architecture is described in Section 3.3.3. Beyond being a baseline, VM-XMorpher is also employed to derive pseudo-labels for the semi-supervised setting.

During initialization, the model was trained on 150 pCT–$CBCT_0$ pairs, using ground-truth labels for the pCT (including the GTV, parotids, larynx, spinal cord, eyes, ears, mandible, esophagus, brainstem, and trachea) and pseudo-labels for $CBCT_0$ generated with VM-XMorpher. In each training cycle, the perception CNN first segmented ROIs in both modalities, and the correspondence CNN (with the same backbone as VM-XMorpher) subsequently registered them. After initialization, training followed the original PC-XMorpher formulation, with the distinction that our setup included approximately 150 labeled CT–$CBCT_0$ pairs, while the remaining CT and CBCT scans were treated as unlabeled. In addition, each of the 150 CT scans was paired with all other CBCTs from the same patient and incorporated into training after initialization. Only $CBCT_0$ was pseudo-labeled, given its closest anatomy to pCT.

Pseudo-label generation: For each cross-validation fold, VM-XMorpher was first trained on all pCT–CBCT pairs, then applied to pCT–$CBCT_0$ to transfer pCT segmentations and produce pseudo-labels used for supervision. CT segmentation typically converged within the first few epochs, whereas $CBCT_0$ segmentation required longer due to lower image quality.

#### Comparative VoxelMorph and PC-Reg-RT

We additionally trained an unsupervised VoxelMorph [73] configured for single-step pCT-to-CBCT registration and PC-Reg-RT [9], using the same pseudo-labels as PC-XMorpher to compensate for the absence of manual CBCT annotations. These models were evaluated alongside XMorpher to identify the most effective framework for subsequent sCT-to-CBCT alignment at test time.

#### Registration Evaluation

Registration evaluation was conducted on the same patients used in the 5-fold cross-validation of the HNC dataset for the synthesis stage. For unsupervised methods, performance was assessed using NCC and SSIM, while target registration error (TRE) was computed for all models.

#### Experimental details

VoxelMorph was trained following the original protocol [73], except with a batch size of 14. XMorpher and PC-Reg-RT were trained for an additional 200 epochs beyond their original configurations. PC-XMorpher and PC-Reg-RT each included an initial pretraining phase of 40 epochs on pCT–$CBCT_0$ pairs with available labels to train the perception CNN for CBCT segmentation, after which training proceeded according to their respective original protocols.

### 4.5.3 MRI/sCT-to-CBCT Registration Evaluation During Testing

To evaluate the complete pipeline from synthesis to registration, a total of 56 patients from the HNC dataset were used for testing. In the synthesis stage, sCT images were generated using the best-performing synthesis method and evaluated using the image similarity and dose metrics. The resulting sCTs, along with their segmentations, were then used in the registration stage. In this stage, the sCTs were deformably aligned to the daily CBCTs using the best-performing registration method, and performance was assessed using NCC and SSIM.

# CHAPTER 5 ARTICLE 1: DIFFUSION-BASED IMAGE SYNTHESIS FOR DEFORMABLE MRI TO CBCT REGISTRATION WITH CROSS-ATTENTION IN HEAD AND NECK RADIOTHERAPY

This article was submitted to the journal Medical Physics on the 22rd October 2025, for peer review.

I contributed to the literature review, methodological design, experimentation, data analysis, and manuscript writing. My contribution is estimated at about 80% of the total work.

**Authors:** Shima Sargordi[1], William T. Le[1], Zeinab Abboud[1], Redha Touati[1], Samuel Kadoury[1,2]

[1] Computer and Software Engineering Department, Polytechnique Montréal, Montréal, QC H3T 1J4, Canada

[2] Centre hospitalier de l'Université de Montréal (CHUM), Montréal, QC H2X 3E4, Canada

Corresponding author: samuel.kadoury@polymtl.ca

**Background:** Magnetic resonance imaging (MRI)-only radiotherapy offers a promising alternative to traditional workflows by eliminating exposure to radiation from planning computed tomography (pCT). However, in adaptive radiotherapy (ART) for head and neck cancer (HNC), the pCT is vital for dose calculation, and accurately aligning it with daily cone-beam CT (CBCT) remains essential yet challenging to properly capture anatomical changes throughout treatment course.

**Purpose:** The goal of this work is to develop an unsupervised two-stage deep learning pipeline for MRI-to-CBCT deformable registration for ART., comprising of: (1) MRI-to-CT synthesis and (2) deformable registration of synthesized CT (sCT) to the daily CBCTs.

**Methods:** In the synthesis stage, following a rigid alignment and image normalization, the framework integrates an improved MRI-to-CT transformer-based denoising diffusion probabilistic model (MC-IDDPM) adopted for its superior performance over GAN-based methods, such as the conditional generative adversarial network (CGAN) and the feature-consistency variant (FCGAN). The synthesized image was then integrated with VM-XMorpher for deformable registration, resulting in an end-to-end MRI-to-CBCT alignment pipeline.

**Results:** From an internal HNC cohort with 338 patients, 282 were used for 5-fold cross-validation, where the MC-IDDPM reduced the the mean absolute error (MAE) by 22.7% and increased the Pearson correlation coefficient (PCC) and peak signal-to-noise ratio (PSNR) by 0.3% and 8.1%, respectively, along with a 1.3% reduction in dose MAE and dose–volume histogram (DVH) difference with test-time augmentation, compared with FCGAN, demon-

strating improved image fidelity and enhanced dosimetric consistency. The FCGAN outperformed the baseline CGAN in all metrics. In the registration stage, VM-XMorpher achieved the highest structural similarity index Measure(SSIM) ($0.92 \pm 0.002$) and the lowest target registration error (TRE) across all methods, with a TRE of $0.67 \pm 0.37$ mm, indicating superior CT-to-CBCT alignment.

**Conclusions:** Our study demonstrated that MC-IDDPM for MRI-to-CT synthesis provides improved anatomical fidelity and enhanced dose accuracy, while VM-XMorpher for deformable registration achieves superior overall performance in sCT-to-CBCT alignment for HNC cases. When the synthesis stage is combined with daily deformable registration, the resulting pipeline establishes a promising framework for MRI-based ART.

## 5.1 Introduction

Adaptive radiotherapy (ART) remains the goal of imaging-based raditherapy cancer treatments [74–76]. It enables precise dose conformation and improves tumor control while reducing irradiation-induced side-effects, by adjusting the treatment plan daily with respect to the patient's anatomical changes. In an increasing number of institutions, acquiring Cone Beam Computed Tomography (CBCT) has become routine as a daily pre-treatment procedure for patient positioning [77]. Despite the physical limitations of CBCT, the information it contains over the course of treatment could enable accurate low-cost plan adaptation by leveraging daily CBCT [78,79], without requiring re-acquisitions of CT, contouring and dosimetric calculations, as well as the entire teams of experts (oncologist, radiologist, medical physicists, *etc.*) associated [80]. In medical image analysis, this task is known as multimodal image registration, which involves aligning images from different modalities (e.g., CT to CBCT) that differ significantly in quality, noise, and anatomical appearance due to daily variations. The goal is to accurately align the original high-quality treatment plan, created on CT, with each daily CBCT [81–85]. While treatment planning is routinely performed on CT due to its high spatial resolution and attenuation coefficient-based imaging, enabling accurate downstream dosimetric calculations [42], there is a growing shift toward Magnetic Resonance Imaging (MRI)-based treatment planning. MRI excels at superior soft tissue contrast and is a non-ionizing, non-invasive imaging modality [86]. With the advent of generative image synthesis techniques [87, 88], MRI-only radiotherapy shows significant promise for clinical use in cancer treatment. Critically, generating synthetic CT (sCT) from MRI eliminates the need for separate CT acquisition, thereby avoiding additional radiation exposure and reducing patient discomfort, while providing CT-equivalent data for downstream dose calculation and registration.

Generative Adversarial Networks (GANs) [28] and conditional GANs (CGANs) [38] are widely used for medical image synthesis tasks like MRI-to-CT translation. Specifically, CGAN models leverage paired data to learn cross-modal mappings, showing promising results for brain MRI-to-CT synthesis. [39,40]. However, head and neck cancer (HNC) synthesis presents additional challenges due to anatomical complexity [41, 42]. While GAN-based approaches using U-Net [6] generators have been applied to head and neck MRI-to-CT synthesis [41, 43, 44], they usually rely on a combination of adversarial loss and pixel-wise losses such as L1 or L2 [43, 45, 46]. This often results in difficulties in preserving fine anatomical structures: a critical limitation [42] for radiotherapy dose calculation.

Denoising Diffusion Probabilistic Models (DDPM) [47] offer improved training stability without the need for adversarial fine-tuning. MRI-to-CT transformer-based improved denoising diffusion probabilistic model (MC-IDDPM) [22] also showed improvement in most image similarity evaluation metrics in MRI-to-CT translation compared to GAN-based approaches, including MC-GAN [49] and MC-CGAN [49], but there is no dosimetric comparison between their proposed method and those GAN-based approaches. Also, their high computational demands for medical image synthesis (reported to be up to 1000 times slower than GANs [22, 89–91]) remain a significant challenge. Although this approach improves synthesis efficiency by reducing the number of generation timesteps, it remains computationally expensive compared to GANs.

In radiotherapy planning, the first step is the delineation of the Planning Target Volumes (PTVs) and Organs at Risk (OARs) on planning CT (pCT), a task typically performed by a radiologist. Automated segmentation algorithms have the potential to further improve clinical workflows by accelerating treatment planning and reducing clinician workload [54]. TotalSegmentator [54], a deep learning toolkit based on nnUNet [55], is trained on 1,204 clinical CT scans and achieved an average Dice similarity coefficient (DSC) of 0.943, covering 104 anatomical structures, including several OARs in the head and neck region. Beyond real CT (rCT), TotalSegmentator can also be applied to sCTs, enabling automatic segmentation of sCTs.

Traditional registration methods rely on iterative optimization techniques [71, 92, 93], which is computationally prohibitive for daily ART [94]. Learning-based approaches offer faster alternatives but face modality constraints.

Unsupervised methods (e.g., VoxelMorph [95]) predict deformations directly from image pairs using similarity losses, enabling real-time performance. Its extension, TransMorph [67], incorporates transformers to capture global context and enhance registration accuracy. Unsupervised models provide an efficient approach for deformable registration, but

they operate without explicit anatomical guidance, which can make the alignment of low-contrast or blurry structures more challenging [96]. Semi-supervised methods, such as Perception-Correspondence Registration with Reverse Teaching (PC-Reg-RT) [96], leverage sparse anatomical labels to improve robustness and mitigate misalignment and distortion of low-contrast structures in few-shot scenarios. It decouples perception (Region of Interest (ROI) identification) and correspondence (alignment) into two Convolutional Neural Networks (CNNs). This separation enables independent optimization: the Perception CNN segments anatomical regions using limited labels, while the Correspondence CNN aligns the ROI to preserve texture and avoid distortion. To handle the few-shot scenario, it introduces Reverse Teaching, where labeled images are aligned with unlabeled ones to generate supervision that teaches the perception network richer anatomical structure and style knowledge. Previous methods, such as VoxelMorph and PC-Reg-RT, relied on a single correspondence CNN that processed concatenated inputs through a shared U-Net backbone, often resulting in distorted feature representations in mixed regions [10]. XMorpher [10] introduces a full transformer architecture that enhances structural correspondence, mitigates feature distortion, and achieves more accurate registration in both unsupervised (VM-XMorpher) and semi-supervised (PC-XMorpher) settings.

Semi-supervised state-of-the-art registration techniques such as PC-Reg-RT [96] and PC-XMorpher [10] have thus far been validated only on unimodal tasks (CT–CT or MRI–MRI), leaving multimodal alignment largely unexplored. This unimodal setting facilitated ROI delineation due to consistent contrast characteristics and the absence of modality-specific artifacts. Furthermore, the anatomical targets were relatively well-defined structures, such as cardiac structures or the cervical vertebrae, rather than the complex structures of OARs in the head and neck region, whose boundaries are considerably more difficult to delineate, particularly in a low-quality modality such as CBCT. In unimodal frameworks, segmentation labels can be directly propagated from the moving to the fixed image, enabling efficient training with consistent anatomical structures. In contrast, CBCTs in our multimodal setting are inherently unlabeled, exhibit lower image quality, and contain artifacts that obscure anatomical boundaries, making structure delineation considerably more difficult than in high-quality CT.

In this study, we propose a unified MRI-based registration framework that can enable ART by integrating two stages: (1) CT generation from diagnostic MRI and (2) deformable registration between CT/sCT and daily CBCT. In the synthesis stage, we adopted the MC-IDDPM fraemwork, which improves structural fidelity and dose estimation accuracy from synthesized CT by incorporating anatomical constraints leveraging domain-specific representations, and establishing it as the primary synthesis approach in our pipeline. It demonstrates

improvements in anatomical coherence compared to standard CGAN models or it's enhanced Feature consistency CGAN (FCGAN). In the deformable registration stage, we adapted the VM-XMorpher model, originally developed for unimodal registration, but in the context of multimodal registration. This two-stage pipeline, consisting of MC-IDDPM–based synthesis and VM-XMorpher–based registration, provides a unified MRI-driven framework for ART that mitigates the limitations of both synthesis and registration in state-of-the-art methods.

## 5.2 Materials and Methods

For the two-stage pipeline—(1) MRI-to-CT synthesis and (2) deformable registration between CT/sCT and daily CBCT, the public SynthRad2023 brain dataset [1] was used primarily to train and validate the synthesis models using image-similarity metrics. An institutional HNC cohort was used for both synthesis and registration tasks within the full MRI-to-CBCT pipeline and, additionally, for the evaluation with segmentation accuracy and dose-based metrics. The following subsections describe the datasets, preprocessing procedures, and the rigid and affine registration steps, followed by a detailed overview of the proposed pipeline, comparative methods, and implementation details. Figure 5.1 depicts the training pipeline of our two-stage framework, and Figure 5.2 illustrates the inference pipeline, in which the sCT produced in stage 1 serves as the input to the second, deformable-registration stage.

### 5.2.1 Datasets and preprocessing

#### SynthRad2023 brain dataset and preprocessing

We first used the publicly available dataset containing 180 paired MRI/CT patients treated with external beam radiotherapy in the brain, to assess the synthesis pipeline. Images were acquired from three centers between 2018 and 2022. The MRI sequences included T1-weighted gradient echo or inversion-prepared turbo field echo (TFE) sequences, with or without Gadolinium contrast agent. The CT scans were acquired without contrast agent, with in-plane resolutions ranging from 0.51 to 1.27 millimeters, slice thicknesses between 1 and 3 millimeters, and energy levels from 80 to 120 kVp peak. The images were cropped to a uniform size of 192×192×164 voxels. Both CT and MRI images were normalized to the range [-1, 1].

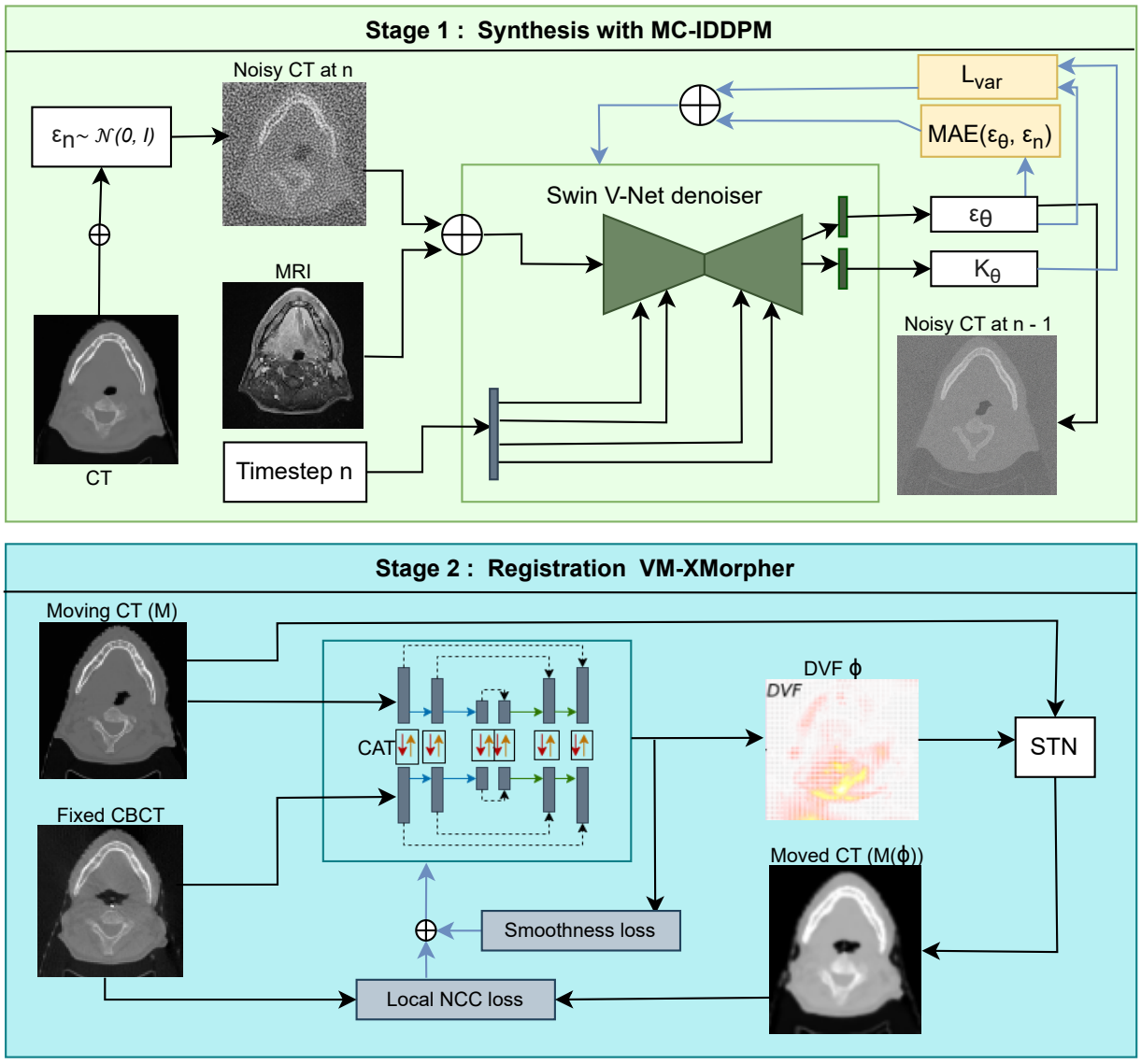


Figure 5.1 Training overview for two stages: *(1) Synthesis* (MC-IDDPM) and *(2) Registration* (VM-XMorpher). In synthesis training stage (Green box above), given MRI, diffusion timestep $n$, and a noisy CT ($x_n$), the model (Swin V-Net denoiser) predicts the noise $\epsilon_\theta$ and the variance coefficient $k_\theta$ and is optimized with $\text{MAE}(\epsilon_\theta, \epsilon_n) + \gamma L_{\text{var}}$. The forward diffusion from $x_0$(clean CT) to $x_n$, in which a clean CT is added (represented by "+") with noise from a Gaussian distribution, is illustrative and not part of the training pipeline; $x_{n-1}$ is shown only to indicate the reverse update and is not used during training. The "Timestep" box represents the sinusoidal embedding that converts the current diffusion step n into a 128-dimensional vector. This embedding is then expanded by a linear (timestep-expanding) layer before being injected into the layers of the Swin V-Net denoiser, allowing each block to know the exact diffusion step and adjust its denoising accordingly. In registration training stage (blue box below), VM-XMorpher, which uses a Cross-Attention Transformer (CAT) blocks, learns a deformation field $\phi$ from CT/CBCT pairs using a local NCC loss (window = 9) with smoothness regularization; the spatial transformer network (STN) applies $\phi$ to the moving CT within the training pipeline.

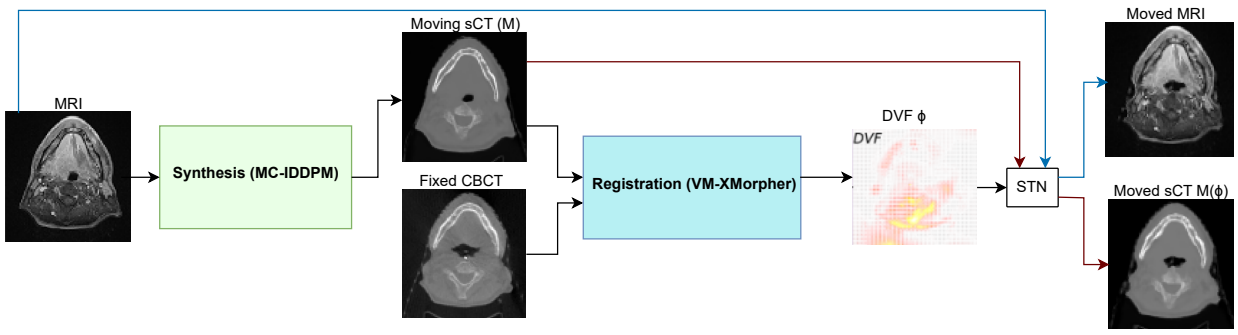


Figure 5.2 Overview of the testing (inference) pipeline: the process performs end-to-end MRI-to-CBCT registration, where MC-IDDPM first synthesizes a CT from the input MRI, and VM-XMorpher subsequently aligns it with the daily CBCT. VM-XMorpher predicts a deformation vector field $\phi$ from the (sCT, CBCT) pair. The spatial transformer network (STN) then applies $\phi$ to the sCT and MRI to generate the warped images.

**Institutional HNC dataset and preprocessing**

A separate retrospective cohort including 338 patients with HNC who underwent curative-intent radiotherapy at the Centre Hospitalier de l'Université de Montréal, as used for synthesis and whole-pipeline evaluation. In this dataset, each patient has one pCT scan (120 kVp, non-contrast) with associated segmentation (e.g., PTVs and OARs) and a radiotherapy plan, as well as one T1-weighted MRI scan and 30–35 CBCT scans. In the synthesis stage, paired MRI/CT images were used and resampled to an isotropic voxel spacing of 1 mm, whereas in the registration stage, CT/CBCT pairs were employed and resampled to 2 mm. For testing, MRI, CT, and CBCT images were all resampled to an isotropic voxel spacing of 1 mm.

CT images were cropped around the tumor volume, and all modalities were normalized using the mean and standard deviation calculated within the head-and-neck region. The resulting volumes after rigid and affine registration (described in the subsequent subsection) were sized 256×256×79 voxels for synthesis stage inputs (MRI/CT pairs) and 128×128×128 voxels for registration stage inputs (CT/CBCT pairs). Following MC-IDDPM [22] in the synthesis stage, images were normalized to [-1, 1], and following XMorpher [10] in the registration stage, images were scaled to [0, 1].

### 5.2.2 Affine pre-alignment for MRI–CT and CT–CBCT pairs

Prior to each fraction of radiotherapy, rigid registration is commonly employed to ensure accurate patient positioning. By aligning the pCT with on-treatment imaging CBCT, the

necessary couch shifts can be determined, allowing the patient's current setup to approximate the position recorded during treatment planning [97].

Before deformable registration, affine alignment was performed for MRI–CT and CT–CBCT pairs in the HNC dataset to ensure initial anatomical correspondence. In contrast, the publicly available SynthRad2023 dataset consists of MRI–CT pairs that were already rigidly registered to correct large inter-modality misalignments, while small anatomical deformations were intentionally left uncorrected. As deformation-corrected CTs are not available in clinical practice, the dataset includes only rigidly aligned images, acknowledging that some deep-learning CT synthesis methods benefit from improved data alignment [1].

Modality pairs (MRI/CT and CT/CBCT) in the HNC dataset were registered for training purposes using Elastix [71] (version 5.2.0) with four resolution levels. For each patient, the pCT, along with its segmentation and dose plan, was first affinely registered to the initial CBCT ($CBCT_0$) using normalized cross-correlation. The MRI was then registered to the aligned pCT using normalized mutual information. The remaining CBCT scans were rigidly registered to $CBCT_0$ using normalized cross-correlation.

Although this affine step is effective for initial alignment, its reliance on global transformations (translation, rotation, scaling, shearing) limits its ability to capture local anatomical variations. Consequently, residual discrepancies may persist in the positioning of targets and OARs, particularly when images are acquired at different times or under varying patient conditions. Moreover, affine methods cannot adequately propagate dose in the presence of non-linear anatomical changes. To overcome these limitations and achieve accurate anatomical correspondence, deformable registration is required [97].

### 5.2.3 Image synthesis with MC-IDDPM

We used the transformer-based improved denoising diffusion probabilistic model (MC-IDDPM) [22] to synthesize CT from MRI. The method consists of a forward diffusion process $q$ that gradually converts a CT from $x_{n-1}$ into Gaussian noise by adding small amounts of noise over many timesteps according to a fixed variance schedule as:

$$q(x_n|x_{n-1}) = \mathcal{N}(x_n;\ \sqrt{1-\beta_n}x_{n-1}, \beta_n\mathbf{I}) \tag{5.1}$$

where $x_n$ is the noisy version at time step $n$, $\beta_n$ is the variance schedule, and $\mathbf{I}$ is the identity matrix. After many steps, the image becomes nearly pure Gaussian noise: $x_N \sim \mathcal{N}(0, \mathbf{I})$.

The reverse diffusion is performed by the trained MC-IDDPM denoiser, conditioned on the

patient's MRI: at each step, it removes a small amount of injected noise from the current noisy CT volume, gradually transforming it back into the clean CT volume that is anatomically consistent with the input MRI.

MC-IDDPM uses a 3D shifted-window transformer VNet (Swin-VNet) denoiser [22] that takes a noised CT volume at timestep $n$, the timestep embedding, and the conditioning MRI. The network predicts (i) the added Gaussian noise and (ii) a learned coefficient for the variance in the reverse step, enabling prediction of both mean and variance of the noise at each step. The shifted-window self-attention in Swin-VNet captures long-range anatomy [22]. The predicted mean can be recovered using the predicted noise:

$$\mu_\theta(x_n, n|Z) = \frac{1}{\sqrt{1-\beta_n}}\left(x_n - \frac{\beta_n}{\sqrt{1-\Sigma_{i=1}^{n}1-\beta_i}}\epsilon_\theta(x_n, n|Z)\right). \tag{5.2}$$

Here, Z is the condition MRI, and $\epsilon_\theta(x_n, n|Z) \sim \mathcal{N}(0, \mathbf{I})$ is the predicted noise by the network. To estimate the variance, the network learns an interpolation factor $k_n(x_n, n|Z)$, and defines:

$$\Sigma_\theta(x_n, n|Z) = \exp(k_n(x_n, n|Z)\log\beta_n + (1-k_n(x_n, n|Z)) \times \log\left(\frac{(1-\prod_{i=1}^{n-1}(1-\beta_i))\beta_n}{1-\prod_{i=1}^{n}(1-\beta_i)}\right) \tag{5.3}$$

where $\beta_n$ is a pre-defined variance corresponding to the diffusion timestep n, and the only unknown parameter $k_\theta$ is obtained by minimizing a variational lower bound loss:

$$\arg\min_{k_\theta} L_{\text{var}} = \arg\min_{k_\theta} \mathcal{L}_{\text{VLB}}(\Sigma_n, \Sigma_\theta(x_n, n|Z)) \tag{5.4}$$

and $\Sigma_n$ is the true variance, and $\Sigma_\theta(x_n, n|Z)$ is the variance predicted by the network.

At inference, a reduced, resampled set of diffusion timesteps is employed to accelerate sampling while maintaining fidelity [22]. Training minimizes a two-term objective:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{mean}} \;+\; \gamma\,\mathcal{L}_{\text{var}}. \tag{5.5}$$

Here, $\mathcal{L}_{\text{mean}}$ is minimized to supervise accurate noise prediction (using an MAE between the true and predicted noise). The scalar $\gamma$ balances the two terms [22].

After training, the estimated mean and variance are used to iteratively denoise the image from Gaussian noise back to a clean image. During inference, 50 timesteps are uniformly sampled from $[1, 1000]$. Starting from pure noise, the reverse process reconstructs the final

sCT aligned with the MRI. The less noisy CT ($x_{n-1}$) can be obtained:

$$x_{n-1} = \mu_\theta(x_n, n|Z) + \sigma_\theta(x_n, n|Z)\,\epsilon \tag{5.6}$$

where $\sigma_\theta$ represents the standard deviation of the reverse diffusion distribution. $\epsilon$ is noise sampled from a Gaussian distribution.

### 5.2.4 Deformable registration with XMorpher

For the deformable registration step with CT/CBCT pairs from the HNC dataset, we adapted both the unsupervised and semi-supervised variants of XMorpher [10] for our multimodal CT-to-CBCT registration pipeline. In addition, as part of the complete MRI-to-CBCT workflow, the sCT generated by the MC-IDDPM model was deformably registered to the daily CBCT to enable MRI-based alignment and evaluation.

In the unsupervised setting, VM-XMorpher [10] was adapted for multimodal CT-to-CBCT registration. The correspondence CNN in VM-XMorpher is a transformer-based registration network that uses dual U-shaped encoder–decoder paths for the fixed and moving images. At each resolution level, the two streams exchange information via Cross-Attention Transformer (CAT) blocks, which establish multi-level semantic correspondences while preserving modality-specific features [10]. The decoder aggregates these fused features to predict a dense deformation vector field (DVF), which is then applied to warp the moving image.

Training followed the original unsupervised formulation: the network is optimized using a local normalized cross-correlation (NCC) loss with a uniform window size of 9 between the warped CT (or sCT) and the fixed CBCT as the similarity loss, combined with a regularization term to encourage smooth deformations. The loss function used for training is defined as:

$$\mathcal{L}(F, M, \phi) = \mathcal{L}_{\text{sim}}(F, M(\phi)) + \lambda \mathcal{L}_{\text{smooth}}(\phi) \tag{5.7}$$

$\lambda$ is a regularization weight that balances the similarity term and the smoothness term. The similarity term encourages alignment between the fixed image $F$ and the warped moving image $M(\phi)$ using the deformation $\phi$. $\mathcal{L}_{\text{smooth}}$ is $\mathcal{L}_2$ regularization on the spatial gradients of the deformation field.

This unsupervised VM-XMorpher served as both a baseline registration method and as the pseudo-label generator for initializing the semi-supervised PC-XMorpher and PC-Reg-RT.

In the semi-supervised setting, we employ PC-XMorpher [10] for multimodal registration. In the initialization stage, we trained PC-XMorpher on 150 pCT–$CBCT_0$ pairs from the HNC dataset using ground-truth labels (including the Gross Tumor Volume (GTV), parotids, larynx, spinal cord, eyes, ears, mandible, esophagus, brainstem, and trachea) for the pCT and VM-XMorpher–derived pseudo-labels for $CBCT_0$. Within each training cycle, the perception CNN segments ROIs in both modalities, followed by the correspondence CNN, which shares the same backbone as VM-XMorpher and registers these ROIs. After initialization, the remainder of the training follows the procedure of the original PC-XMorpher [10], with the modification that our setting included approximately 150 labeled CT–$CBCT_0$ pairs, while the remaining CT and CBCT images were treated as unlabeled. In addition, each of these 150 CT scans was also paired with the other CBCTs from the same patient and incorporated into training after the initialization stage.

Pseudo-label generation: To enable PC-XMorpher in the multimodal setting, pseudo-labels were generated for the first-day CBCT ($CBCT_0$) using VM-XMorpher. VM-XMorpher was trained on all pCT–CBCT pairs within each training fold. After training, it was applied to the paired pCT–$CBCT_0$ in that fold to estimate the DVFs and transfer the pCT segmentations onto $CBCT_0$, thereby generating pseudo-labels. This step was restricted to $CBCT_0$, which is anatomically most similar to the pCT and thus provides the most reliable label transfer. These pseudo-labels, together with the ground truth pCT labels, were then used as auxiliary supervision for training PC-XMorpher in the semi-supervised setting. Pseudo-labels were required since the perception CNN was unable to segment CBCT images without supervision. CT segmentation converged in the early epochs, whereas $CBCT_0$ segmentation required approximately 30 additional training epochs to converge.

The semi-supervised model was optimized using the same NCC loss (window size of 9) and smoothness regularization as in the VM-XMorpher formulation (see equation 5.7), with the addition of a cross-entropy loss between the predicted and reference labels.

### 5.2.5 Comparative methods

#### Image synthesis with CGAN and FCGAN

This section first outlines the baseline CGAN architecture employed and then explains the feature-consistency loss incorporated to enhance the FCGAN framework. So, FCGAN is similar to CGAN, with only an additional loss term introduced into the generator loss.

**Network architecture** The baseline CGAN was inspired by the Pix2Pix framework [38], which has been adapted for MRI-to-CT synthesis. The generator adopts a U-Net [6] with skip connections. The encoder has four blocks: block 1 applies two $3 \times 3$ convolutions (stride 1, padding 1) with ReLU activation and no downsampling; blocks 2–4 each begin with a $2 \times 2$ max-pooling layer (stride 2), followed by two $3 \times 3$ convolutions (stride 1, padding 1) with ReLU. The channel widths are 32, 64, 128, and 256 for blocks 1–4, respectively. The bottleneck applies an additional $2 \times 2$ max-pooling operation, then two $3 \times 3$ convolutions to expand to 512 channels (stride 1, padding 1, ReLU), and upsamples back to 256 channels using a transposed convolution (kernel size 1, stride 2, output padding 1). The decoder also has four blocks; at each level, encoder features are concatenated channel-wise with the decoder features, followed by two $3 \times 3$ convolutions (stride 1, padding 1) with ReLU. Upsampling between decoder levels is performed with transposed convolutions (kernel size 1, stride 2, output padding 1). No normalization or dropout layers are used in the generator. A final $1 \times 1$ convolution maps 32 channels to the output channel, followed by a Tanh activation to produce the synthesized CT image.

The discriminator follows a PatchGAN design inspired by Pix2Pix [38], producing a patch-wise real/fake prediction map. It consists of five convolutional layers with a kernel size of $4 \times 4$ and padding of 1. The first four layers use a stride of 2, while the final layer uses a stride of 1. Each layer is followed by instance normalization (except the first and last) and a LeakyReLU activation ($\alpha = 0.2$).

**Loss functions** The baseline CGAN generator is optimized using a combination of Binary Cross-Entropy (BCE) loss and $\mathcal{L}_1$ loss between the predicted CT and the ground-truth CT. As for the FCGAN, we introduce an additional constraint to the baseline CGAN. The FCGAN generator loss is therefore defined as:

$$\mathcal{L}_{\text{gen}} = \tfrac{1}{3}\Big[\mathcal{L}_{\text{BCE}}(D(G(x), x), \mathbf{1}) + \|G(x) - y\|_1 + \|f_1 - f_2\|_1\Big] \tag{5.8}$$

where $\|\cdot\|_1$ denotes the mean absolute error, $x$ is the input condition MRI, $G(x)$ is the predicted CT (or sCT) image from the generator $G$, $D$ is the discriminator, $y$ is the ground truth CT image or rCT, and $\mathbf{1}$ is a matrix of ones. $f_1$ represents features extracted from the input MRI at the final layer of the generator's encoder, to which a 2D convolution layer is applied to adjust the output dimensionality to match the discriminator's feature map output, and $f_2$ represents features extracted by the discriminator from the rCT image. This early-stage alignment guides the generator to internalize the structural characteristics of rCTs before image synthesis, leading to greater preservation of anatomical structures in

the synthesized CT.

The discriminator loss for both CGAN, FCGAN is defined as:

$$\mathcal{L}_{\text{disc}} = \frac{1}{2}\Big[\|D(G(x), x) - \mathbf{0}\|_1 + \|D(y, x) - \mathbf{1}\|_1\Big] \tag{5.9}$$

where $D(G(x), x)$ is the discriminator output for the input pair of sCT and MRI, $D(y, x)$ is the discriminator output for the input pair of rCT and MRI, and $\mathbf{0}$ and $\mathbf{1}$ are matrices of zeros and ones, respectively.

**Deformable registration with VoxelMorph and PC-Reg-RT**

We further adapted unsupervised VoxelMorph [73] and semi-supervised PC-Reg-RT [94].

We adapted the VoxelMorph framework by Le et al. [73], using a single-step registration of the pCT to each daily CBCT, instead of the original two-stage strategy [73]. Furthermore, we adapted PC-Reg-RT by He et al. [96], and to address the absence of manual annotations in CBCTs, we leveraged the same pseudo-labels generated for PC-XMorpher. The best-performing method is then used for sCT-to-CBCT alignment during testing.

### 5.2.6 Implementation details

To evaluate the synthesis module, we employed a 5-fold cross-validation (CV) for the SynthRad2023 dataset, ensuring that each fold contained approximately equal proportions of data from all three sites.

Of the 338 patients in the HNC dataset, 56 were reserved for testing, while the remaining 282 were used for 5-fold CV. For both the synthesis and registration stages, the same patient splits were used for the 5-fold CV and the test cohort.

FCGAN and the baseline CGAN were implemented in PyTorch 2.0.1 with Python 3.10.14 and trained on an NVIDIA GeForce RTX 2080 Ti GPU (11 GB memory). The models were optimized using Adam with an initial learning rate of $1 \times 10^{-5}$. A step-based learning rate scheduler was employed, reducing the learning rate by a factor of 0.8 every 100 epochs. The number of training epochs ranged from 100 to 240, determined by early stopping on the validation set based on the best image similarity metrics, including mean absolute error (MAE), peak signal-to-noise ratio (PSNR), Pearson correlation coefficient (PCC), and multi-scale structural similarity index Measure (MS-SSIM). For MC-IDDPM, we followed the same implementation as in the original work [22]. While the original study trained and inferred on 64×64×4 sub-volumes, we used larger patches (192×192×4 for the SynthRad2023 brain

dataset and 256×256×4 for the HNC dataset) as the smaller patch size yielded suboptimal results.

For the registration methods, VoxelMorph was trained following the original strategy [73], except for using a batch size of 14. VM-XMorpher, PC-XMorpher, and PC-Reg-RT were each trained for an additional 200 epochs beyond their original configurations (originally 400 epochs for VM-XMorpher [10] and 200 epochs for both PC-XMorpher [10] and PC-Reg-RT [96]). Both PC-XMorpher and PC-Reg-RT underwent an initial training phase using only pCT $CBCT_0$ pairs with labels for 40 epochs, during which the perception CNN learned to predict CBCT segmentations. The subsequent training followed the original protocol [10, 96] of 200 epochs, extended by an additional 200 epochs.

## 5.3 Performance evaluation

We present a separate evaluation for each stage of the proposed framework, along with a final evaluation of the entire pipeline using the best-performing model from each stage.

### 5.3.1 sCT evaluation

To comprehensively assess the sCTs generated by all models, we evaluated both image similarity metrics and generated dose metrics.

#### sCT image similarity evaluation

To compare MC-IDDPM [22] with CGAN and FCGAN, we employed several image similarity metrics, including MAE, PSNR, PCC, and MS-SSIM (with an evaluation scale of 5). Higher PSNR, PCC, and SSIM values indicate better sCT quality. Using these metrics, we compared MC-IDDPM with the baseline CGAN and FCGAN on the SynthRad2023 brain dataset. All three synthesis methods were also trained on the HNC dataset. The model configuration for the final pipeline was selected based on its superior overall performance across image similarity and dose metrics presented in Section 5.4.

For the HNC dataset, we additionally report the Dice similarity coefficient (DSC) between the structures segmented on the sCT and those on the reference CT, including the spinal cord, parotid glands, and skull, as well as dose-related metrics (details are provided in Section 5.3.1). TotalSegmentator [54] was used to provide head and neck segmentations, including the parotid glands, spinal cord, and skull (covering the mandible), which we used for evaluating the sCT images. PTVs or Other OAR segmentations are either unavailable or not

publicly provided in TotalSegmentator. It is used to segment sCT images in the HNC dataset, providing labels for DSC computation in 5-fold CV and in the test-set. These evaluations enable a comprehensive assessment of both voxel-level similarity and structural consistency across different anatomical regions.

#### sCT dose evaluation

To clinically evaluate the sCT images, we performed a dosimetric assessment using the Swin UNETR++ deep learning–based dose prediction model proposed by Wang et al. [56], which was trained and validated on the public OpenKBP HNC dataset [57], leading to a Dose Score of 2.84 and a dose–volume histogram (DVH) Score of 1.76, and subsequently applied it to our in-house HNC validation and test set. For each patient, dose distributions were predicted using both the rCT and the corresponding sCT. This approach ensured that all patients were evaluated under a standardized protocol, as treatment plans were generated with identical configurations and constraints, thereby allowing a direct comparison between the synthetic and real CTs. Evaluation metrics included the MAE of the dose and the DVH difference, both computed relative to the ground-truth dose calculated on the rCT. Test-time augmentation (TTA) was implemented by applying multiple augmentations (e.g., flips, rotations, scaling) to the input during inference and averaging the predictions to improve robustness [56].

The dose MAE was computed over the entire dose mask covering the head and neck region. DVH differences were reported in two ways: (1) an overall average DVH score, obtained by aggregating the errors across all three target volumes (PTV70, PTV63, and PTV56 represent high-, intermediate-, and low-risk tumor regions, at prescribed doses of 70, 63, and 56 Gy, respectively) as well as the OARs (including the spinal cord, parotid glands, brainstem, larynx, and mandible), and (2) per-structure DVH metrics calculated individually for each target volume and each OAR. This evaluation compares sCTs generated by CGAN, FCGAN, and MC-IDDPM. Dosimetric evaluation was performed with real radiotherapy structures because the dose calculation framework [56] requires complete sets of PTVs and OARs. Some of these structures, including all PTVs, were not available in the TotalSegmentator predictions [54]. For dose calculations, all required real structures were aligned to the sCTs using an affine transformation followed by a B-spline registration using Elastix [71], to ensure more accurate alignment with the sCTs.

### 5.3.2 CT to CBCT registration evaluation

To evaluate the initial registration accuracy on the HNC dataset, all unsupervised methods were assessed using masked normalized cross-correlation (NCC) and masked MS-SSIM

(computed across five scales). Target registration error (TRE) was additionally measured on 5 patients per fold, based on three manually annotated landmarks placed on clinically relevant structures (the left parotid, right parotid, and spinal cord) across all models. CT-to-CBCT registration was performed during validation using real image pairs, while during testing, sCTs generated from MRI were used instead of real CTs to evaluate MRI-to-CBCT alignment in adaptive radiotherapy.

### 5.3.3 sCT to CBCT registration evaluation using test set

To register MRI to daily CBCTs in the HNC test set, we first generated sCTs from the corresponding MRI scans using the proposed MC-IDDPM. The synthesis performance at test time was assessed using image similarity metrics and dose metrics. These sCTs were then registered to the daily CBCTs using the best-performing registration model trained on the entire set of CT–CBCT pairs. The resulting DVFs were used to warp the segmentations, and the MRI to each daily CBCT. Registration performance was evaluated using masked NCC and masked MS-SSIM.

## 5.4 Results

In this section, we first present the image similarity results for CT synthesis using CGAN, FCGAN, and MC-IDDPM on both the SynthRad2023 brain dataset and the HNC dataset, followed by the corresponding dosimetric evaluation conducted on the HNC dataset only. We then report the performance of deformable CT-to-CBCT registration methods. Finally, we present the synthesis test results and their subsequent use in the sCT-to-CBCT registration task using the entire pipeline.

### 5.4.1 MRI-to-CT synthesis

Figure 5.3 and Figure 5.4 display sCT images generated from the SynthRad2023 brain dataset [1] and the HNC dataset, respectively. More specifically, Figure 5.4 also presents the dose distributions predicted from the real CT and the corresponding sCT, along with the associated dose difference map.

The quantitative evaluation of CGAN, FCGAN, and MC-IDDPM for MRI-to-CT synthesis is detailed in the following subsection.

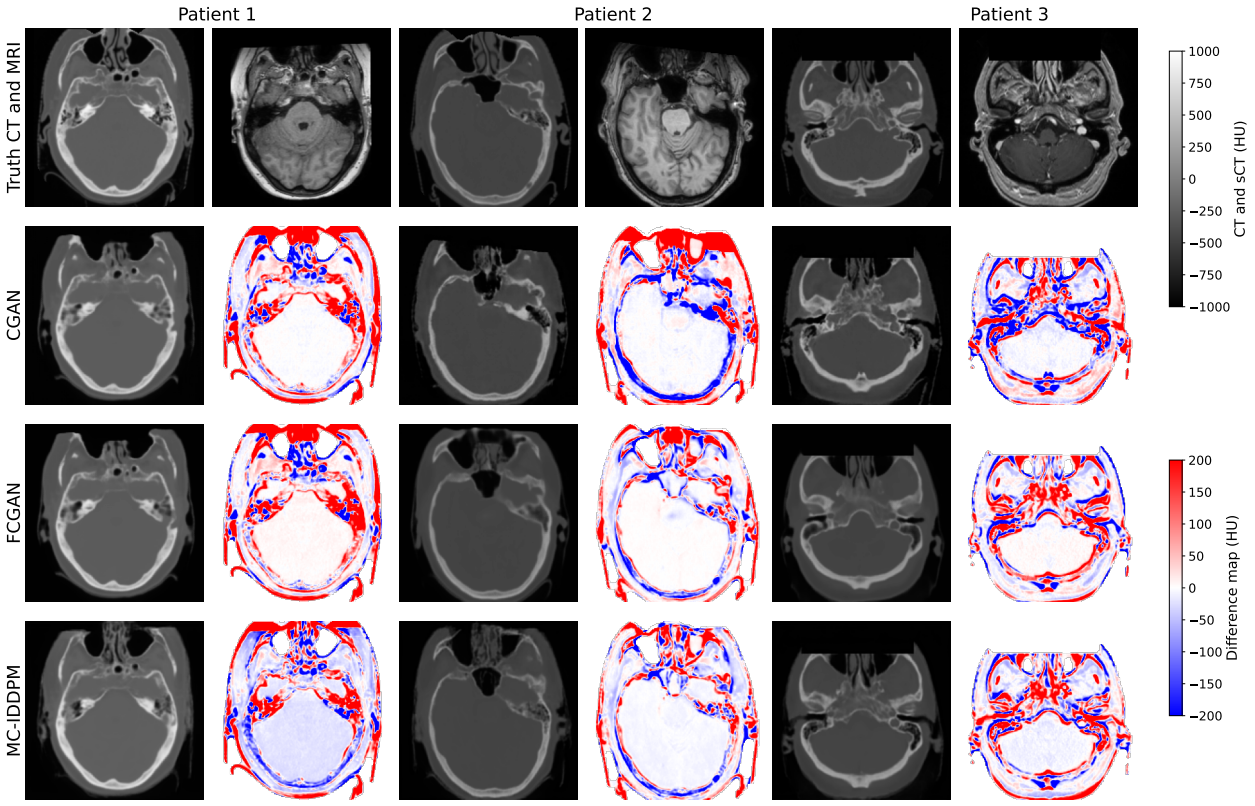


Figure 5.3 SynthRad2023 brain synthesized CT (sCT) comparison. sCT images were generated from the SynthRad2023 Task 1 brain validation set. The first row displays ground-truth CT images (left) and the corresponding input MRIs (right). Rows two to four show the sCT outputs from CGAN (row 2), FCGAN (row 3), and MC-IDDPM (row 4). Columns one, three, and five present the sCTs, while columns two, four, and six show the corresponding difference maps relative to the ground-truth CTs.

### Image similarity results

The results on the SynthRad2023 [1] brain dataset are summarized in Table 5.1. For computing these similarity metrics, sCT and corresponding rCT voxel intensities were normalized to [0,1]. MC-IDDPM achieved the lowest MAE ($0.039 \pm 0.006$) and the highest PCC ($0.959 \pm 0.002$), PSNR ($24.04 \pm 0.8$), and SSIM ($0.91 \pm 0.002$), outperforming the GAN-based methods across all image similarity metrics. Amongst the GAN-based models, FCGAN showed improved performance over CGAN with a lower MAE ($0.040 \pm 0.002 < 0.044 \pm 0.003$), higher PSNR ($22.71 \pm 0.1 > 21.67 \pm 0.7$), higher PCC ($0.949 \pm 0.002 > 0.928 \pm 0.012$), and higher SSIM ($0.88 \pm 0.005 > 0.86 \pm 0.018$). Overall, MC-IDDPM demonstrated superior accuracy and structural fidelity compared with both CGAN and FCGAN on the SynthRad2023 brain dataset.

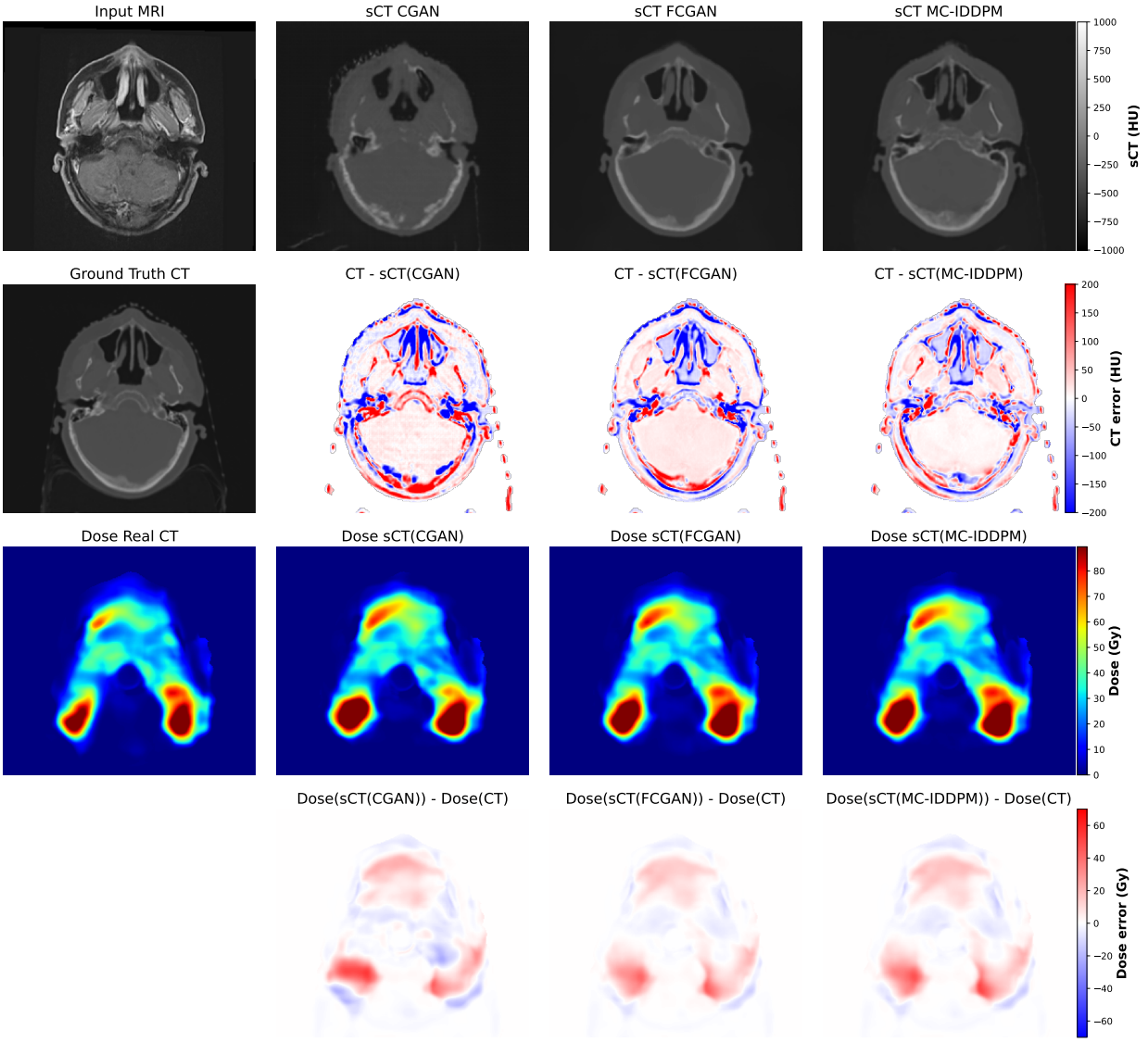


Figure 5.4 MRI-to-sCT and dose comparison using HNC validation set. The first row shows the input MRI and the synthesized CT (sCT) images generated by CGAN, FCGAN, and MC-IDDPM. The second row presents the corresponding ground-truth CT and the difference map for each method. The third row displays the real dose distribution computed on the ground-truth CT alongside the dose maps calculated on each sCT. The last row shows the dose-difference maps for the same models.

We further evaluated all synthesis methods on the HNC dataset in Table 5.2. MC-IDDPM achieved the lowest MAE ($0.034 \pm 0.002$) and the highest PCC ($0.939 \pm 0.002$) and PSNR ($25.96 \pm 0.2$ dB), demonstrating superior image fidelity compared to both GAN-based methods. FCGAN outperformed the baseline CGAN, achieving a lower MAE ($0.044 \pm 0.003$ vs. $0.062 \pm 0.011$) and higher PCC ($0.936 \pm 0.003$ vs. $0.930 \pm 0.008$) and PSNR ($24.02 \pm 0.4$

Table 5.1 MRI-to-CT synthesis image similarity metrics on the SynthRad2023 dataset using 5-fold cross-validation.

| Model | MAE | PCC | PSNR (dB) | SSIM |
|---|---|---|---|---|
| CGAN | 0.044 ± 0.003 | 0.928 ± 0.012 | 21.67 ± 0.7 | 0.86 ± 0.018 |
| FCGAN | 0.040 ± 0.002 | 0.949 ± 0.002 | 22.71 ± 0.1 | 0.88 ± 0.005 |
| **MC-IDDPM** | **0.039 ± 0.006** | **0.959 ± 0.002** | **24.04 ± 0.8** | **0.91 ± 0.002** |

Bold values indicate the best result in each column.

Table 5.2 MRI-to-CT synthesis image similarity metrics on the HNC dataset using 5-fold cross-validation comparing MC-IDDPM with CGAN and FCGAN.

| Model | MAE | PCC | PSNR(dB) | SSIM |
|---|---|---|---|---|
| CGAN | 0.062 ± 0.011 | 0.930 ± 0.008 | 22.37 ± 0.8 | 0.90 ± 0.011 |
| FCGAN | 0.044 ± 0.003 | 0.936 ± 0.003 | 24.02 ± 0.4 | **0.91 ± 0.001** |
| **MC-IDDPM** | **0.034 ± 0.002** | **0.939 ± 0.002** | **25.96 ± 0.2** | 0.91 ± 0.006 |

Bold values indicate the best result in each column.

dB vs. $22.37 \pm 0.8$ dB), indicating better pixel-level accuracy and noise reduction. MC-IDDPM and FCGAN achieved identical SSIM values ($0.91 \pm 0.001$ vs $0.91 \pm 0.006$), reflecting strong structural preservation; however, MC-IDDPM consistently provided improved voxel-wise similarity, outperforming CGAN across all metrics and FCGAN in MAE, PCC, and PSNR.

Figure 5.5 presents DSC scores on the HNC dataset for the spinal cord, parotids, and skull. Segmentations were obtained using TotalSegmentator [54] applied to either the sCT or the corresponding rCT, and evaluated against the clinical radiotherapy ground-truth (GT) segmentation when available. For example, the DSC between the spinal cord GT segmentation and the spinal cord segmentation obtained from TotalSegmentator with sCT is $0.66 \pm 0.01$ for FCGAN, $0.65 \pm 0.01$ for MC-IDDPM, and $0.62 \pm 0.02$ for CGAN. As a reference, the DSC between the spinal cord GT segmentation and the corresponding rCT segmentation from TotalSegmentator is $0.69 \pm 0.01$, representing the expected upper bound of agreement for spinal cord, as TotalSegmentator [54] was trained and validated on CT scans with minimal preprocessing (only resampling to isotropic resolution), it is inherently more reliable when applied to unpreprocessed images. Furthermore, direct comparisons between TotalSegmentator-derived segmentations from rCT and sCT yielded high DSC values. For the skull, MC-IDDPM reached $0.87 \pm 0.010$, higher than FCGAN ($0.86 \pm 0.001$) and CGAN ($0.85 \pm 0.010$). These results confirm that the sCTs preserve anatomical structure sufficiently for consistent automated segmentation with all methods.

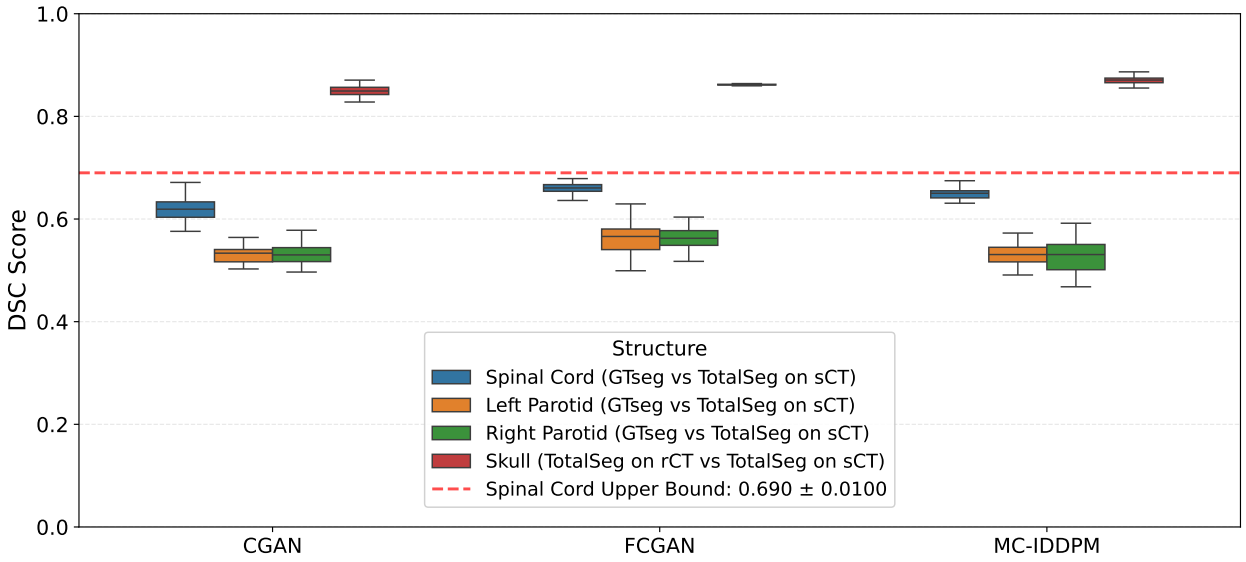


Figure 5.5 Dice similarity coefficient (DSC) values for selected anatomical structures using CGAN, FCGAN, and MC-IDDPM. Ground-truth RT segmentations (GTseg) are compared with TotalSegmentator (TotalSeg) predictions on sCT. A red dashed line indicates the DSC of spinal cord segmentations between GT and TotalSeg on real CT (rCT) for the spinal cord.

For the parotid glands, MC-IDDPM achieved DSC values of $0.53 \pm 0.02$ for the left and $0.53 \pm 0.03$ for the right, compared to FCGAN ($0.56 \pm 0.03$ and $0.56 \pm 0.02$, respectively), which were higher. The CGAN model achieved a DSC of $0.53 \pm 0.02$ for both the left and right parotid glands. The overall DSC values for the parotids were lower than those for other structures, reflecting the greater difficulty of segmenting these glands due to their small size and higher contouring variability. Overall, FCGAN shows better performance on soft tissues (spinal cord, parotids), while MC-IDDPM performs better on bony structures (skull) across all three methods.

In summary, on SynthRad2023, MC-IDDPM outperformed FCGAN with a 2.5% reduction in MAE and an increase of 1.1%, 5.9%, and 3.4% in PCC, PSNR, and SSIM, respectively. On the HNC cohort, improvements were more pronounced, with a 22.7% lower MAE, and 0.3% and 8.1% gains in PCC and PSNR. On SynthRad2023, FCGAN achieved a 9.1% reduction in MAE and increase of 2.3%, 4.8%, and 2.3% in PCC, PSNR, and SSIM, respectively; on the HNC dataset, the gains were 29.0%, 0.6%, 7.4%, and 1.1%. Relative to FCGAN, MC-IDDPM improves skull Dice by 1.2%, while being lower on spinal cord by 1.5% and lower on parotid left/right by 5.4% each. Compared with CGAN, FCGAN increases Dice for the

spinal cord by 6.5%, skull by 1.2%, and parotids by 5.7% each.

Table 5.3 Dose metrics comparing predicted doses from sCTs on the HNC dataset. TTA denotes test-time augmentation.

| Model | MAE | MAE (TTA) | DVH | DVH (TTA) |
|---|---|---|---|---|
| CGAN | 3.063 ± 0.20 | 1.573 ± 0.08 | 3.035 ± 0.48 | 0.782 ± 0.06 |
| FCGAN | 2.950 ± 0.08 | 1.567 ± 0.05 | 2.774 ± 0.25 | 0.779 ± 0.04 |
| **MC-IDDPM** | **2.134 ± 0.07** | **1.547 ± 0.09** | **1.007 ± 0.09** | **0.769 ± 0.08** |

Bold values indicate the best result in each column.

### Dosimetric results

Table 5.3 summarizes the dose evaluation metrics on the HNC dataset. For all dose metrics, MC-IDDPM outperformed CGAN and FCGAN, achieving a dose MAE of $1.547 \pm 0.09$ with TTA. Without TTA, MC-IDDPM also yielded lower dose MAE (2.134±0.07) than CGAN and FCGAN. MC-IDDPM also showed superior performance in DVH difference, obtaining 0.769± 0.08 with TTA, indicating closer agreement with rCT-based dose calculations. Without TTA, it also achieved the lowest DVH difference of $1.007 \pm 0.09$. Note that TTA is only applied to dose metrics calculation.

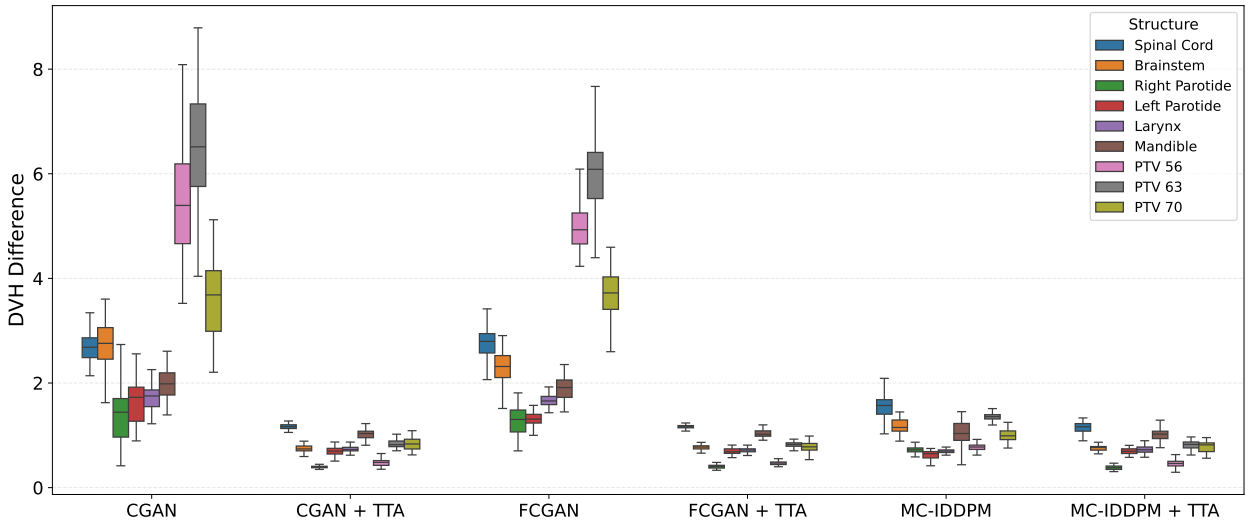


Figure 5.6 Boxplots of DVH differences for various anatomical structures using three synthesis models (CGAN, FCGAN, and MC-IDDPM), with and without test-time augmentation (TTA). Each box shows the distribution across five-fold cross-validation; structures include organs at risk (OARs) and planning target volumes (PTVs).

Figure 5.6 compares DVH differences in doses predicted on sCTs from CGAN, FCGAN, and MC-IDDPM, evaluated with and without TTA. Applying TTA substantially reduced DVH discrepancies for the GAN-based models (CGAN and FCGAN), where large variations across OARs and PTVs were markedly reduced. For MC-IDDPM, TTA provided additional but smaller improvements, as the baseline performance was already consistent across structures. Overall, these results highlight that TTA significantly enhances the dosimetric accuracy of GAN-based methods across all structures while providing smaller improvements for most structures in MC-IDDPM.

The results reported below correspond to the lowest DVH differences obtained for each structure across all methods (whether with or without TTA). MC-IDDPM achieved the lowest DVH differences across most structures, with TTA yielding the lowest differences for the spinal cord ($1.158 \pm 0.10$), brainstem ($0.742 \pm 0.06$), mandible ($1.015 \pm 0.11$), and PTV70 ($0.784 \pm 0.11$). Without TTA, it achieved the lowest DVH differences for the left parotid ($0.619 \pm 0.09$) and larynx ($0.702 \pm 0.04$). FCGAN with TTA performed better in PTV56 ($0.462 \pm 0.05$ ) and PTV63 ($0.807 \pm 0.06$), whereas MC-IDDPM obtained DVH differences of $0.463 \pm 0.06$ and $0.813 \pm 0.09$ for PTV56 and PTV63, respectively. These results demonstrate that MC-IDDPM provides superior dose preservation in most OARs and in PTV70, as well as improved dosimetric consistency compared to GAN-based methods. Moreover, its voxel-wise dose accuracy (dose MAE) was lower than that achieved by CGAN or FCGAN.

As a result, compared to FCGAN, MC-IDDPM reduced dose MAE by 27.7% and DVH difference by 63.7% without TTA, and achieved an additional 1.28% reduction for both MAE(TTA) and DVH(TTA) with TTA. Compared to CGAN, FCGAN reduced dose MAE by 3.7% and DVH difference by 8.6% without TTA, and by 0.4% for both metrics with TTA.

### 5.4.2 sCT-to-CBCT registration result

Table 5.4 summarizes the quantitative registration performance of VM-XMorpher compared with all other methods, including the B-spline registration with NiftyReg [59], the unsupervised VoxelMorph, and the semi-supervised approaches PC-XMorpher and PC-Reg-RT. VM-XMorpher achieved the highest SSIM of $0.92 \pm 0.002$ and the lowest TRE $0.67 \pm 0.37$ mm among all methods, indicating superior structural correspondence and landmark alignment accuracy between the warped CTs and daily CBCTs. While VoxelMorph obtained a slightly higher NCC ($0.91 \pm 0.002$) compared with VM-XMorpher ($0.88 \pm 0.001$), its TRE ($0.80 \pm 0.55$ mm) was higher, indicating less accurate alignment compared to VM-XMorpher. B-spline registration with NiftyReg (TRE = $1.14 \pm 0.86$ mm) and PC-Reg-RT (TRE = $1.07 \pm 0.69$ mm) exhibited higher TRE values compared with other methods, indicating less accurate spatial

alignment. Overall, VM-XMorpher achieved the best performance in the similarity metric SSIM and in geometric alignment, demonstrating superior accuracy across most evaluation metrics and better alignment between the patient pCT and its daily CBCTs.

VM-XMorpher reduced landmark error by 16.2% relative to VoxelMorph, 41.2% relative to B-spline, 37.4% relative to PC-Reg-RT, and 22.1% relative to PC-XMorpher, and improved SSIM by 1.1%, 3.4%, and 8.2% compared to VoxelMorph, B-spline, and affine/rigid initialization, respectively.

Table 5.4 Registration performance on the HNC dataset using 5-fold cross-validation. Metrics reported include NCC (normalized cross-correlation), SSIM (structural similarity index measure), and TRE (target registration error in mm).

| Model | NCC | SSIM | TRE (mm) |
|---|---|---|---|
| Affine/rigid init. (Elastix) | 0.76 ± 0.003 | 0.85 ± 0.003 | 2.14 ± 1.79 |
| B-spline [59] | 0.80 ± 0.003 | 0.89 ± 0.003 | 1.14 ± 0.86 |
| **VM-XMorpher** | 0.88 ± 0.001 | **0.92 ± 0.002** | **0.67 ± 0.37** |
| VoxelMorph | **0.91 ± 0.002** | 0.91 ± 0.002 | 0.80 ± 0.55 |
| PC-XMorpher | Semi | | 0.86 ± 0.30 |
| PC-Reg-RT | Supervised | | 1.07 ± 0.69 |

The table highlights the best-performing method in bold.

### 5.4.3 MRI-to-CT synthesis and sCT-to-CBCT registration results

Using the MC-IDDPM–generated sCTs for the deformable registration stage, as it achieved superior similarity metrics in Tables 5.1 and 5.2 and yielded lower dose MAE and DVH difference than FCGAN and CGAN. In the HNC synthesis test set with MC-IDDPM, we obtained an MAE of $0.040 \pm 0.018$, PCC of $0.939 \pm 0.031$, PSNR of $25.01 \pm 2.6$, SSIM of $0.92 \pm 0.024$ and DSC for spinal cord ($0.67 \pm 0.086$), parotid left ($0.49 \pm 0.16$), parotid right ($0.53 \pm 0.12$), and skull ($0.87 \pm 0.08$). MC-IDDPM obtained dose MAE of $1.53 \pm 0.42$ and DVH difference of $0.758 \pm 0.50$ with TTA. VM-XMorpher obtained NCC of $0.92 \pm 0.016$, SSIM of $0.92 \pm 0.019$ on deformed sCT–CBCT pairs from the test set.

## 5.5 Discussion

This study presents a two-stage, MRI-driven pipeline for daily guidance in head-and-neck radiotherapy, consisting of diffusion-based MRI-to-CT synthesis followed by deformable sCT-to-CBCT registration. Across two datasets and evaluation metrics, MC-IDDPM achieved the most accurate synthesized images, while VM-XMorpher provided the most reliable geometric

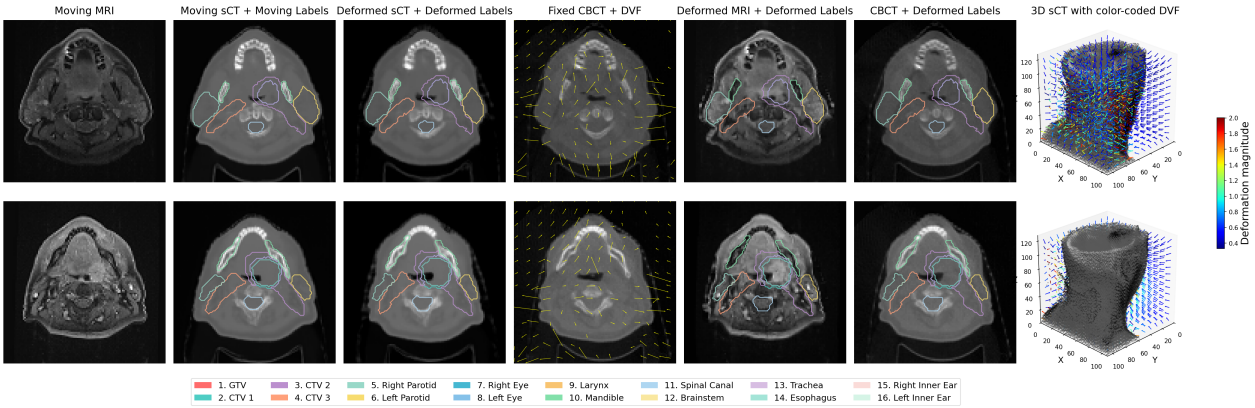


Figure 5.7 VM-XMorpher MRI-to-CBCT registration on test set. Each row shows: (1) Input moving MRI, (2) The generated moving sCT from the MRI and its segmentations, (3) moved sCT and segmentations, (4) fixed CBCT with DVF, (5) moved MRI and segmentations, (6) fixed CBCT with deformed segmentations, and (7) the 3D deformed sCT and DVF. Moving sCT segmentations are radiotherapy reference segmentations that are deformed to the sCT using affine and B-spline transformations. The listed segmentations are those deformed, and not all of them appear in this visualization slice. Cilinical target volume (CTV) 3, CTV 2, and CTV 1 correspond to PTV56, PTV63, and PTV70, respectively.

alignment, making their combination a promising foundation for MRI-based ART. The first synthesis stage eliminates the need for an additional CT, thereby reducing acquisition time, cost, and patient discomfort, while the registration stage enables plan adaptation to daily anatomical changes, ultimately improving treatment precision and patient quality of life.

On SynthRad2023, MC-IDDPM outperformed FCGAN with a reduction in MAE and an increase in PCC, PSNR, and SSIM. On the HNC cohort, improvements were more pronounced, with more MAE reduction, and gains in PCC and PSNR. These findings indicate that MC-IDDPM preserves voxel-wise intensity and structural consistency more effectively, which is essential for accurate dose prediction. The feature-consistency loss in FCGAN produced consistent gains over the baseline CGAN, improving structural fidelity. On SynthRad2023, FCGAN reduced MAE and increased PCC, PSNR, and SSIM compared to CGAN. On the HNC dataset, FCGAN also outperformed CGAN. These results indicate that aligning high-level MRI encoder features with rCT discriminator features stabilizes anatomical representation, reduces pixel-wise error, and improves structural similarity in the challenging HNC anatomy, and the generated sCTs better preserve anatomical fidelity in FCGAN compared to CGAN. Relative to FCGAN, MC-IDDPM improves skull Dice while being lower on spinal cord by and lower on parotid left/right. Compared with CGAN, FCGAN increases Dice for all structures. Overall, FCGAN performs better in soft tissues (spinal cord and parotids),

whereas MC-IDDPM performs better in bony structures.

Image similarity alone is not sufficient for treatment planning. The dose must remain stable under modality substitution. By coupling similarity metrics with dose prediction on identical plan constraints, we show sCT dosimetry tracks rCT more closely with MC-IDDPM than with GANs. Using a standardized learned dose predictor [56], MC-IDDPM produced the lowest dose errors among all methods. Compared to FCGAN, MC-IDDPM reduced dose MAE and DVH difference with and without TTA. TTA substantially benefited the GANs, while MC-IDDPM remained comparatively robust even without it, which is consistent with its stronger voxel-wise fidelity. Practically, these dose results show that the sCTs generated by MC-IDDPM more closely match rCT dose distributions, preserving voxel-wise intensity and attenuation properties essential for dose calculation. Compared to CGAN, FCGAN reduced dose MAE and DVH difference with and without TTA. While TTA improved the CGAN/FCGAN dosimetric errors and tightened agreement with rCT-based dose, enhancing OAR sparing and target coverage, MC-IDDPM still provided the most reliable dosimetric performance across all methods.

Inference with CGAN and FCGAN can be completed within minutes, whereas MC-IDDPM requires several hours and at least three times more GPU memory, not accounting for its substantially longer training time. This computational burden remains the primary limitation of diffusion-based synthesis for clinical deployment.

In the registration stage, VM-XMorpher was employed for deformable alignment of CT/sCT to daily CBCTs. All CT/CBCT pairs were resampled to 2-mm isotropic resolution. Thus, one voxel corresponds to 2 mm when interpreting landmark errors. VM-XMorpher achieved the highest SSIM and the lowest TRE, indicating that registration uncertainty remains below the image sampling and is not the dominant source of geometric error during contour propagation or dose accumulation. VoxelMorph obtained a higher NCC compared to VM-XMorpher, which is consistent with its global NCC loss, emphasizing overall correlation. In contrast, VM-XMorpher optimizes a local NCC objective, emphasizing structural correspondence within OARs and PTVs at the expense of global intensity alignment. This result aligns with prior findings that local or ROI-aware similarity functions yield better anatomical correspondence despite slightly lower global metrics. VM-XMorpher reduced landmark error compared to VoxelMorph, B-spline, PC-Reg-RT, and PC-XMorpher, and improved SSIM compared to VoxelMorph, B-spline, and affine/rigid initialization, respectively. Such behavior is clinically favorable, as maintaining precise OAR and target alignment is more critical for adaptive decision-making than maximizing global correlation.

The reported TRE values were derived from three OAR landmarks, while SSIM and NCC

provide more robust volumetric assessments. Collectively, these results confirm that unsupervised methods such as VM-XMorpher and VoxelMorph outperform semi-supervised variants. Although pseudo-labels from VM-XMorpher on $CBCT_0$ enabled multimodal training, PC-XMorpher and PC-Reg-RT failed to surpass VM-XMorpher, primarily due to CBCT artifacts and noise that impede the perception network from reliable segmentation. VM-XMorpher's cross-attention fusion, which operates without CBCT labels, proved more robust under these conditions. Pseudo-label generation remains critical but also highlights the limitations of current semi-supervised methods on noisy modalities. Future work should incorporate advanced segmentation to improve label reliability and generalization. In a clinical MRI-based ART workflow, expert manual segmentations are routinely available and could also directly strengthen such models.

In summary, diffusion-based sCT generation has improved dosimetric, voxel-wise, and structural fidelity over GAN baselines, while registration with cross-attention provides the most anatomically reliable deformations for HNC. Together, these advances support a promising, clinically feasible pathway toward MRI-based ART.

## 5.6 Conclusion

This work presents a unified framework for MRI-to-CBCT image registration that integrates diffusion-based MRI-to-CT synthesis with transformer-based deformable registration to advance MRI-only ART in HNC. The MC-IDDPM model demonstrated the highest image and dosimetric fidelity, achieving lower voxel-wise errors and improved PSNR and PCC values, while consistently reducing dose discrepancies in both MAE and DVH metrics compared with GAN-based approaches. The inclusion of feature-consistency learning in FCGAN enhanced structural realism over the baseline CGAN, yet the diffusion-based MC-IDDPM model provided greater robustness and cross-dataset generalization. Clinically, the significant improvement in DVH deviations for key organs at risk, particularly the brainstem, spinal cord, mandible, parotid glands, and larynx, underscore the potential of MC-IDDPM for dose preservation in MRI-only workflows.

For the deformable registration component, VM-XMorpher achieved the highest SSIM and the lowest TRE among all tested methods, indicating accurate structure-preserving alignment between sCT and CBCT. Together, the combination of diffusion-based synthesis and transformer-based registration enables consistent geometric and dosimetric correspondence across modalities, supporting a pathway toward fully MRI-driven ART.

**Acknowledgments**

This work was supported by MITACS (Canada).

**Conflict of Interest Statement**

No conflict of interest to declare.

**Ethical approval**

All procedures involving human participants complied with the ethical standards of the institutional and national research committees and with the 1964 Declaration of Helsinki and its later amendments, or comparable ethical standards. The other authors have no relevant conflicts of interest to disclose.

# CHAPTER 6 GENERAL DISCUSSION

This study investigates the feasibility of an MRI-based workflow for adaptive radiotherapy in head and neck cancer using a two-stage deep learning pipeline. The first stage performs MRI-to-CT synthesis to enable segmentation and dose calculation, while the second stage applies deformable registration of CT/sCT to daily CBCT scans. In this framework, synthetic CT volumes are generated from MRI and subsequently aligned with daily CBCTs to estimate deformation vector fields (DVFs), allowing spatial transformation of the sCT or MRI across treatment fractions. By replacing real CT acquisition with sCT, the workflow eliminates the need for additional CT imaging, thereby reducing treatment preparation time, cost, and patient discomfort. The proposed approach was designed to ensure anatomical fidelity, preserve dosimetric accuracy, and achieve accurate registration of organs at risk. This work offers a distinct clinical contribution by conducting the first unified assessment of diffusion-based MRI-to-CT synthesis together with transformer-based CT-to-CBCT deformable registration on an HNC dataset. The findings show that MRI-derived sCTs can support segmentation, dose evaluation, and deformable alignment, highlighting their potential as a promising alternative to CT in the ART workflow. This integrated evaluation goes beyond previous studies that examine synthesis and registration independently and establishes a foundation for reducing reliance on repeated CT imaging while maintaining clinically required accuracy, highlighting the promising feasibility of an MRI-only ART workflow.

## 6.1 Summary of the Synthesis Stage

In the synthesis stage, both GAN-based and diffusion-based models were investigated. The initial baseline was a conditional Generative Adversarial Network (CGAN), which combined pixel-wise and adversarial losses for MRI-to-CT generation. Building on this framework, an enhanced version, termed FCGAN, was developed by incorporating a feature-consistency constraint that enforced alignment between high-level MRI encoder representations and discriminator features extracted from real CT images. This additional term guided the generator to internalize anatomical characteristics prior to synthesis, thereby improving structural preservation in the resulting sCTs. On the public SynthRad2023 brain dataset, FCGAN achieved lower MAE and higher PSNR, PCC, and SSIM values compared with CGAN, demonstrating improved voxel-level similarity and structural accuracy. When evaluated on the HNC dataset, FCGAN also yielded consistent gains across SSIM, PCC, and PSNR while further reducing MAE. Dice score analysis confirmed this trend, showing higher overlap between the

synthesized and real CT structures with FCGAN, indicating enhanced anatomical fidelity compared to the baseline CGAN. This improvement directly mitigates a key limitation of conventional GAN-based approaches, which often struggle to preserve fine anatomical structures due to their reliance on simple L1 or L2 losses. In addition, the inclusion of the feature-consistency constraint led to more stable training dynamics, as FCGAN continued to improve with additional epochs, whereas the baseline CGAN exhibited less stability over extended training. MC-IDDPM demonstrated superior performance across all image similarity metrics on the SynthRad2023 dataset. On the HNC cohort, it further improved MAE, PCC, and PSNR compared to FCGAN, confirming its ability to generate more quantitatively accurate sCTs. In terms of structural agreement, MC-IDDPM achieved higher Dice scores in bony regions, indicating better preservation of dense anatomical structures such as the skull. However, slightly lower Dice scores were observed in soft-tissue regions, including the parotid glands and spinal cord, suggesting that its representation of soft-tissue contrast was less precise than that of FCGAN. Image similarity alone does not guarantee clinical validity for radiotherapy planning, as the stability of dose distribution under modality substitution is more critical. By integrating image similarity metrics with dose prediction under identical constraints, the results demonstrate that the dosimetric accuracy of sCTs generated by MC-IDDPM more closely follows that of real CTs compared to GAN-based models. MC-IDDPM consistently achieved the lowest dose errors across all evaluation metrics, including the MAE of the dose distribution and the dose–volume histogram (DVH) difference. Test time augmentation notably improved the performance of GAN-based methods, whereas MC-IDDPM maintained robust dosimetric consistency even without it, reflecting its higher voxel-wise fidelity and stable intensity mapping. From a clinical standpoint, these findings indicate that the sCTs generated by MC-IDDPM more robustly maintains real CT dose distributions, ensuring reliable preservation of voxel-wise intensity and attenuation properties essential for dose calculation and treatment planning.

### 6.1.1 Limitations of the Synthesis Stage

First, MC-IDDPM synthesis is computationally expensive, with both training and inference times substantially longer than those of GAN-based models. While GAN-based inference can be completed within minutes for all cases in the validation or test set, inference with MC-IDDPM requires several hours, which limits its practicality for repeated generation. Second, in this study the dose evaluation of synthesized CTs relied on a learned dose prediction model [56] rather than a physics-based recalculation of the treatment plan. While this approach allows fast and consistent comparisons, it can introduce bias from the prediction model itself. As a result, the comparison may reflect not only differences between

the real and synthetic images but also errors from the prediction model itself, which can make the evaluation less specific to the quality of the sCT. Third, to advance toward a fully automatic and clinically promising pipeline, complete PTV/OAR segmentations at test time from a validated learned model would be required. At test time, we used complete clinical RT contours rather than automatic segmentations from a learned model; consequently, we did not evaluate a fully automatic replanning workflow. Fourth, training a diffusion model is comparatively more straightforward, as its optimization process relies solely on minimizing a well-defined loss function, which inherently reduces the risk of underfitting or overfitting. In contrast, GAN-based methods, particularly the baseline CGAN, presented substantial challenges during training, including instability. Moreover, adversarial losses in GANs are often less informative, making it difficult to determine convergence or select the optimal training epoch. In practice, model selection depended heavily on validation metrics, which revealed a trade-off between minimizing voxel-wise errors (e.g., MAE) and enhancing structural similarity (e.g., SSIM). This inherent instability and competing optimization objectives made GAN-based models more difficult to train and less predictable compared to diffusion-based approaches.

## 6.2 Summary of Deformable Registration Stage

The second stage of the framework focused on deformable alignment between CT and CBCT images, using both unsupervised and semi-supervised approaches adapted to the multimodal CT–CBCT setting. Among all evaluated models, VM-XMorpher demonstrated more precise structural correspondence. Its dual encoder–decoder design, coupled with cross-attention modules, enables mutual feature exchange between modalities, which reflects tight anatomical correspondence in critical regions. The use of a localized normalized cross-correlation (NCC) loss guided the model toward region-specific alignment, yielding the highest SSIM and the lowest target registration error (TRE) across all configurations.

In contrast, VoxelMorph, which optimizes a global NCC loss, achieved higher global intensity correlation but lower structural agreement. This discrepancy arises from the underlying optimization objectives of the model. global similarity objectives promote overall alignment but often overlook finer regional details, whereas local similarity constraints emphasize precise correspondence within organs at risk (OARs) and target volumes. From a clinical standpoint, the latter is more desirable, as accurate localization within OARs is critical for dose accumulation and adaptive replanning.

Semi-supervised variants such as PC-XMorpher and PC-Reg-RT were adapted by generating pseudo-labels on $CBCT_0$ to guide the perception network. While this strategy allowed

supervision in the absence of CBCTs ground-truth labels, its performance was hindered by the limited accuracy of pseudo-segmentations and by the inherent difficulty of delineating structures in noisy, low-contrast CBCTs. The perception CNN, originally designed for high-contrast modalities, struggled to segment low-contrast or irregularly shaped regions such as the (Gross Tumor Volume) GTV, which can appear in arbitrary locations, as well as small or complex OARs. Consequently, supervision based on pseudo-labels propagated errors through training, leading to weaker correspondence and higher TRE than their unsupervised methods.

Overall, the findings highlight the robustness of unsupervised cross-attention registration in multimodal scenarios. Despite the lack of segmentation supervision, VM-XMorpher better captured spatial correspondence and maintained geometric precision, making it the preferred choice for integration into the end-to-end MRI-to-CBCT pipeline.

### 6.2.1 Evaluating Deformation Smoothness and Statistical Tests

We quantified the proportion of voxels with a non-positive Jacobian determinant as an indicator of local folding. Across the HNC dataset, PC-XMorpher and PC-Reg-RT produced the smoothest fields ($0.002\% \pm 0.001\%$ and $0.003\% \pm 0.001\%$), followed by VoxelMorph ($0.018\% \pm 0.002\%$). VM-XMorpher showed higher folding rates ($0.281\% \pm 0.143\%$) and B-spline interpolation yielded $0.197\% \pm 0.010\%$, indicating that the methods with anatomical guidance provide the most regular deformations, while VM-XMorpher exhibits reduced smoothness but superior alignment accuracy (highest SSIM and lowest TRE).

### 6.2.2 Limitations of the Registration Stage

Stage 2 is constrained by the absence of true CBCT labels, which makes the evaluation of methods more difficult and less accurate. The perception CNN used in semi-supervised methods is not very accurate for CBCT segmentation, particularly for the GTV or low-contrast OARs. Pseudo-labels derived from deformed planning CT provide a practical way to make the models trainable, but their limited accuracy also reduce segmentation effectiveness.

The absence of ground-truth CBCT segmentations also limits the evaluation framework. Current assessments depend primarily on intensity-based metrics and landmark errors, which, while informative, do not fully capture regional misalignments within OARs and target volumes. If accurate segmentation masks were available for CBCTs across fractions, the registration performance could be evaluated using Dice similarity coefficients between warped and reference structures, providing a more comprehensive and anatomically meaningful measure of spatial accuracy than landmark-based or voxel-wise intensity metrics alone. Such seg-

mentation availability would also enable consistent training supervision, facilitating models that jointly optimize structural correspondence and geometric consistency throughout the adaptive workflow.

### 6.3 Summary of the HNC dataset Pre-Processing and limitations

Comprehensive pre-processing was required to ensure that the two stages of the pipeline, MRI-to-CT synthesis and CT/sCT-to-CBCT registration, could operate coherently within a unified framework. The planning CT was first affinely registered to the initial CBCT ($CBCT_0$), and all other CBCTs were rigidly registered to the first CBCT to account for geometric differences between planning and on-treatment imaging. The MRI was subsequently aligned to the affinely registered CT, establishing a common spatial reference across all modalities prior to synthesis and deformable registration. This sequential alignment ensured spatial consistency throughout the workflow but introduced practical challenges arising from differences in the field of view (FOV).

In particular, CBCT volumes typically cover a smaller anatomical region than pCT, leading to the exclusion of CT slices outside the CBCT coverage during affine registration. Consequently, the MRI, aligned to the cropped CT, underwent additional trimming. Since the MRI acquisitions in this HNC cohort already contained a limited number of slices (approximately 75), the cumulative cropping further reduced the available anatomical content compared to the SynthRad2023 brain dataset, which features a wider FOV. Reduced contextual information limits both synthesis fidelity.

Each model required modality-specific intensity normalization tailored to its architecture. While synthesis and registration networks operated within different dynamic ranges, all data were scaled consistently within each stage to ensure stable training. For tasks such as segmentation and dose prediction, preserving the physical meaning of Hounsfield Units (HU) was critical. Therefore, all synthesized CT volumes were converted back to their HU range before processing with TotalSegmentator [54] or the Swin-UNETR++ dose predictor [56]. However, minor discrepancies in voxel intensity remained due to cumulative interpolation and scaling steps during resampling, making exact HU recovery across the full volume infeasible.

To minimize such inconsistencies, future implementations would benefit from a unified preprocessing protocol encompassing voxel spacing, FOV harmonization, and standardized intensity scaling across all modalities. Establishing a single consistent normalization and resampling policy that synthesis, registration, segmentation, and dose prediction operate on consistent inputs can reduce bias.

## 6.4 Future work

Future work should focus on enhancing both the supervision and automation components of the pipeline to improve robustness. For the registration stage, semi-supervised strategies could benefit from a stronger perception backbone capable of reliable CBCT segmentation, thereby providing more accurate supervisory signals.

Extending the framework toward fully automatic segmentation of all PTVs and OARs, on both sCTs for dose calculation and daily CBCTs for registration, would make the MRI-based ART workflow more automatic.

A promising direction involves generating reliable CBCT segmentations to establish Dice-based evaluation of registration accuracy, replacing intensity-based or landmark-based metrics with anatomically grounded measures. Additionally, expert-drawn manual segmentations for CBCTs can be leveraged. Finally, adopting a unified pre-processing and normalization standard across synthesis, registration, and dose modules would mitigate inter-stage inconsistencies and facilitate seamless end-to-end deployment in adaptive radiotherapy.

# CHAPTER 7 CONCLUSION

Radiotherapy is a standard treatment for head-and-neck cancer (HNC) and relies on multimodal imaging. MRI informs target delineation and treatment planning; CT provides electron-density information for dose calculation; and CBCT enables daily monitoring of anatomical change. Adaptive radiotherapy (ART) aims to maintain target coverage while sparing organs at risk by accounting for day-to-day variation over a multi-week course. Within this context, deep-learning methods provide data-driven mechanisms to model these changes and support ART. An MRI-based ART paradigm is to plan with MRI while recovering CT-equivalent properties for dose computation and aligning to daily CBCT. Implementing this paradigm requires two tightly coupled capabilities: robust MRI-to-CT synthesis that preserves dose-critical anatomy, and deformable registration that remains reliable on CBCT despite modality gaps and image degradations.

This thesis presents a unified, end-to-end pipeline that integrates MRI-to-CT synthesis with CT/sCT-to-CBCT deformable registration for adaptive HNC radiotherapy, assessed using image, dose, and registration measures. In Stage 1, both adversarial and diffusion approaches were examined for MRI-to-CT synthesis, and the diffusion model generated CTs (MC-IDDPM) were adopted for the subsequent registration stage for its superiority in image and dose metrics.

In Stage 2, we evaluated both unsupervised and semi-supervised registration methods for CT–to-CBCT alignmnet. VM-XMorpher consistently achieved the best accuracy and was therefore selected for the final pipeline. Semi-supervised variants did not provide additional benefits in our setting due to the limited reliability of CBCT delineations, constrained by the perception network's inaccuracy in segmenting CBCT images and the inherent imprecision of pseudo-labels.

End-to-end, the combination of diffusion-based synthesis and transformer-based registration yielded a robust pipeline for deforming sCT and MRI to daily CBCTs. Paired with a standardized learned dose predictor, the sCTs demonstrated fidelity sufficient for downstream dose evaluation and registration. Taken together, these results support the feasibility of an MRI-based adaptive workflow that reduces reliance on routine pCT acquisition while preserving the capacity for daily assessment and potential adaptation.

In conclusion, this work advances MRI-based ART by demonstrating that top-performing synthesis and registration components can be integrated into a coherent pipeline aligned with clinical endpoints. By employing dose-aware evaluation for synthesis and landmark ac-

curacy for registration, eliminating the need for a separate CT scan and better accounting for anatomical changes in the patient. This can reduce patient discomfort by limiting additional imaging and lower toxicity through more accurate treatment delivery.

# REFERENCES


[1] A. Thummerer *et al.*, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical physics*, vol. 50, no. 7, pp. 4664–4674, 2023.

[2] "Head and neck cancer treatments, causes & symptoms," 2024. [Online]. Available: https://nhcancerclinics.com/cancer-types/head-and-neck-cancer/

[3] B. Ajtai, E. Lindzen, and C. M. Joseph, "Neuroimaging structural imaging: Magnetic resonance imaging, computed tomography," *Bradley's Neurology in Clinical Practice. 6th Edition, Elsevier Saunders*, 2012.

[4] M. Precision, "Computed tomography (ct) basic and principle," 2018, tech Tips column, updated 2025-05-20. [Online]. Available: https://www.matsusada.com/column/words-ct.html

[5] R. Singh, "Decoding cnns: A beginner's guide to convolutional neural networks and their applications," 2024.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *MICCAI*, 2015.

[7] P. Isola *et al.*, "Image-to-image translation with conditional adversarial networks," *CVPR*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1611.07004

[8] G. Balakrishnan *et al.*, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018.

[9] Y. He *et al.*, "Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching," *IEEE Journal of Biomedical and Health Informatics*, 2021.

[10] J. Shi *et al.*, "Xmorpher: Full transformer for deformable medical image registration via cross attention," *MICCAI*, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.07349

[11] G. Henique, "Deep learning for risk stratification in adaptive radiotherapy of head and neck cancers," Master's thesis, Department of Biomedical Engineering, Polytechnique Montreal, 2025.

[12] W. T. Le *et al.*, "Prediction of head and neck radiotherapy toxicity using a deformable 3d cnn on longitudinal daily cbct acquisitions," in *IEEE 20th ISBI*, 2023.

[13] M. Alefsen de Boisredon d'Assier, "Weakly supervised deep learning methods for cross-modality tumor segmentation in medical imaging," Master's thesis, Institute of Biomedical Engineering, Polytechnique Montreal, Montreal, QC, 2023.

[14] M. Sargordi, "Training neural networks to perform structured prediction task," Master's thesis, Department of Computer and Software Engineering, Polytechnique Montreal, 2024, page 6.

[15] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015.

[16] H. Chandarana *et al.*, "Emerging role of mri in radiation therapy," *Journal of Magnetic Resonance Imaging*, 2018.

[17] V. Khoo *et al.*, "Comparison of mri with ct for the radiotherapy planning of prostate cancer: A feasibility study," *The British journal of radiology*, 1999.

[18] N. Chowdhury *et al.*, "Concurrent segmentation of the prostate on mri and ct via linked statistical shape models for radiotherapy planning," *Medical Physics*, vol. 39, no. 4, pp. 2214–2228, 2012.

[19] G. C. Pereira, M. Traughber, and R. F. Muzic, "The role of imaging in radiation therapy planning: past, present, and future," *BioMed Research International*, 2014.

[20] M. Boulanger *et al.*, "Deep learning methods to generate synthetic ct from mri in radiotherapy: a literature review," *Physica Medica*, vol. 89, pp. 265–281, 2021.

[21] P. Decazes *et al.*, "Trimodality pet/ct/mri and radiotherapy: a mini-review," *Frontiers in Oncology*, vol. 10, p. 614008, 2021.

[22] S. Pan *et al.*, "Synthetic ct generation from mri using 3d transformer-based denoising diffusion model," *Medical Physics*, 2024. [Online]. Available: https://doi.org/10.1002/mp.16847

[23] C. M. Rank *et al.*, "Mri-based treatment plan simulation and adaptation for ion radiotherapy using a classification-based approach," *Radiation Oncology*, 2013.

[24] V. Keereman *et al.*, "Mri-based attenuation correction for pet/mri using ultrashort echo time sequences," *Journal of Nuclear Medicine*, 2010.

[25] T. Stanescu *et al.*, "A study on the magnetic resonance imaging (mri)-based radiation treatment planning of intracranial lesions," *Physics in Medicine and Biology*, 2008.

[26] A. Johansson *et al.*, "Improved quality of computed tomography substitute derived from magnetic resonance (mr) data by incorporation of spatial information–potential application for mr-only radiotherapy and attenuation correction in positron emission tomography," *Acta Oncologica*, 2013.

[27] S. Kazemifar *et al.*, "Dosimetric evaluation of synthetic ct generated with gans for mri-only proton therapy treatment planning of brain tumors," *Journal of Applied Clinical Medical Physics*, 2020.

[28] I. J. Goodfellow *et al.*, "Generative adversarial networks," *NeurIPS*, 2014. [Online]. Available: https://doi.org/10.48550/arXiv.1406.2661

[29] K. He *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[30] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[31] D. Pathak *et al.*, "Context encoders: Feature learning by inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [Online]. Available: https://doi.org/10.48550/arXiv.1604.07379

[32] J. Gui *et al.*, "A review on generative adversarial networks: Algorithms, theory, and applications," *IEEE Transactions on Knowledge and Data Engineering*, 2021. [Online]. Available: https://doi.org/10.48550/arXiv.2001.06937

[33] M. Wiatrak and S. V. Albrecht, "Stabilizing generative adversarial network training: A survey," *arXiv preprint*, 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1910.00927

[34] N. Kodali *et al.*, "On convergence and stability of GANs," *arXiv preprint arXiv:1705.07215*, 2017.

[35] H. Ge *et al.*, "Fictitious GAN: Training GANs with historical models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[36] T. Miyato *et al.*, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[37] T. Motwani and M. Parmar, "A novel framework for selection of GANs for an application," *arXiv preprint arXiv:2002.08641*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2002.08641

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

[39] A. Aljohani and N. Alharbe, "Generating synthetic images for healthcare with novel deep pix2pix gan," *Electronics*, 2022.

[40] A. Ranjan, D. Lalwani, and R. Misra, "Gan for synthesizing ct from t2-weighted mri data towards mr-guided radiation treatment," *Magnetic Resonance Materials in Physics, Biology and Medicine*, vol. 35, no. 3, pp. 449–457, 2022.

[41] P. Klages *et al.*, "Patch-based generative adversarial neural network models for head and neck mr-only planning," *Medical Physics*, 2020.

[42] R. Touati, W. T. Le, and S. Kadoury, "A feature invariant generative adversarial network for head and neck mri/ct image synthesis," *Physics in Medicine and Biology*, 2021.

[43] M. Qi and et al., "Multi-sequence mr image-based synthetic ct generation using a generative adversarial network for head and neck mri-only radiotherapy," *Medical Physics*, 2020.

[44] A. M. Dinkla *et al.*, "Dosimetric evaluation of synthetic ct for head and neck radiotherapy generated by a patch-based three-dimensional convolutional neural network," *Medical Physics*, 2019.

[45] D. Nie *et al.*, "Medical image synthesis with context-aware generative adversarial networks," *MICCAI*, 2017.

[46] J. M. Wolterink *et al.*, "Deep mr to ct synthesis using unpaired data," in *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI).* Springer, 2017.

[47] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2006.11239

[48] T. Che *et al.*, "Your gan is secretly an energy-based model and you should use discriminator driven latent sampling," *Advances in Neural Information Processing Systems*, 2020.

[49] Y. Lei *et al.*, "Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks," *Medical Physics*, 2019.

[50] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021.

[51] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[52] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, 2021.

[53] E. M. C. Huijben *et al.*, "Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report," *Medical Image Analysis*, 2024.

[54] J. Wasserthal *et al.*, "Totalsegmentator: Robust segmentation of 104 anatomical structures in ct images," *Radiology AI*, 2023.

[55] F. Isensee *et al.*, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," *arXiv preprint arXiv:1809.10486*, 2018.

[56] K. Wang, H. S. Tan, and R. Mcbeth, "Swin unetr++: Advancing transformer-based dense dose prediction towards fully automated radiation oncology treatments," *arXiv preprint arXiv:2311.06572*, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.06572

[57] A. Babier *et al.*, "Openkbp: The open-access knowledge-based planning grand challenge and dataset," *Medical Physics*, 2021. [Online]. Available: https://doi.org/10.1002/mp.14845

[58] M. Chen *et al.*, "Image registration: Fundamentals and recent advances based on deep learning," in *Machine Learning for Brain Disorders*. Springer, 2023.

[59] M. Modat *et al.*, "Fast free-form deformation using graphics processing units," *Computer methods and programs in biomedicine*, vol. 98, no. 3, pp. 278–284, 2010.

[60] J.-P. Thirion, "Image matching as a diffusion process: an analogy with maxwell's demons," *Medical image analysis*, vol. 2, no. 3, pp. 243–260, 1998.

[61] M. F. Beg *et al.*, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International journal of computer vision*, 2005.

[62] H. Sokooti *et al.*, "Nonrigid image registration using multi-scale 3d convolutional neural networks," *MICCAI*, 2017.

[63] M.-M. Rohé *et al.*, "Svf-net: learning deformable image registration using shape matching," *MICCAI*, 2017.

[64] B. D. De Vos *et al.*, "End-to-end unsupervised deformable image registration with a convolutional neural network," in *International Workshop on Deep Learning in Medical Image Analysis*. Springer, 2017.

[65] A. Dalca *et al.*, "Learning conditional deformable templates with convolutional networks," *Advances in neural information processing systems*, 2019.

[66] M. P. Heinrich *et al.*, "Towards realtime multimodal fusion for image-guided interventions using self-similarities," *MICCAI*, 2013.

[67] J. Chen *et al.*, "Transmorph: Transformer for unsupervised medical image registration," *Medical Image Analysis*, 2022.

[68] J. Chen *et al.*, "A survey on deep learning in medical image registration: New technologies, uncertainty, evaluation metrics, and beyond," *Medical Image Analysis*, 2025.

[69] A. Reithmeir *et al.*, "From model based to learned regularization in medical image registration: A comprehensive review," *Medical Image Analysis*, 2025.

[70] L. S. Bosma *et al.*, "Tools and recommendations for commissioning and quality assurance of deformable image registration in radiotherapy," *Physics and Imaging in Radiation Oncology*, 2024.

[71] S. Klein *et al.*, "Elastix: a toolbox for intensity-based medical image registration," *IEEE transactions on medical imaging*, vol. 29, no. 1, pp. 196–205, 2009.

[72] F. Pérez-García, R. Sparks, and S. Ourselin, "Torchio: a python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning," *Computer methods and programs in biomedicine*, vol. 208, p. 106236, 2021.

[73] W. T. Le *et al.*, "Comparing 3D deformations between longitudinal daily CBCT acquisitions using CNN for head and neck radiotherapy toxicity prediction," *arXiv preprint arXiv:2303.03965*, 2023.

[74] Q. J. Wu *et al.*, "Adaptive radiation therapy: technical components and clinical applications," *The cancer journal*, vol. 17, no. 3, pp. 182–189, 2011.

[75] S. Lim-Reinders *et al.*, "Online adaptive radiation therapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 99, no. 4, pp. 994–1003, 2017.

[76] O. M. Dona Lemus *et al.*, "Adaptive radiotherapy: next-generation radiotherapy," *Cancers*, vol. 16, no. 6, p. 1206, 2024.

[77] D. A. Jaffray *et al.*, "Flat-panel cone-beam computed tomography for image-guided radiation therapy," *International Journal of Radiation Oncology* Biology* Physics*, vol. 53, no. 5, pp. 1337–1349, 2002.

[78] D. N. Stanley *et al.*, "A roadmap for implementation of kv-cbct online adaptive radiation therapy and initial first year experiences," *Journal of applied clinical medical physics*, vol. 24, no. 7, p. e13961, 2023.

[79] E. Lavrova *et al.*, "Adaptive radiation therapy: a review of ct-based techniques," *Radiology: Imaging Cancer*, vol. 5, no. 4, p. e230011, 2023.

[80] A. T. Price *et al.*, "Initial clinical experience building a dual CT- and MR-guided adaptive radiotherapy program," *Clinical and Translational Radiation Oncology*, vol. 42, p. 100661, 2023.

[81] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[82] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable medical image registration: A survey," *IEEE Transactions on Medical Imaging*, vol. 32, no. 7, pp. 1153–1190, 2013.

[83] M. A. Viergever *et al.*, "A survey of medical image registration–under review," pp. 140–144, 2016.

[84] G. Haskins, U. Kruger, and P. Yan, "Deep learning in medical image registration: a survey," *Machine Vision and Applications*, vol. 31, no. 1, p. 8, 2020.

[85] S. Bharati *et al.*, "Deep learning for medical image registration: A comprehensive review," *arXiv preprint arXiv:2204.11341*, 2022.

[86] J. M. Edmund and T. Nyholm, "A review of substitute ct generation for mri-only radiation therapy," *Radiation Oncology*, vol. 12, no. 1, p. 28, 2017.

[87] R. Touati, W. T. Le, and S. Kadoury, "Multi-planar dual adversarial network based on dynamic 3d features for mri-ct head and neck image synthesis," *Physics in Medicine & Biology*, vol. 69, no. 15, p. 155012, 2024.

[88] R. Touati and S. Kadoury, "A least square generative network based on invariant contrastive feature pair learning for multimodal mr image synthesis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 6, pp. 971–979, 2023.

[89] J. Peng *et al.*, "CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model," *Medical physics*, vol. 51, no. 3, pp. 1847–1859, 2024.

[90] Q. Lyu and G. Wang, "Conversion between ct and mri images using diffusion and score-matching models," *arXiv preprint arXiv:2209.12104*, 2022.

[91] J. Wolleb *et al.*, "Diffusion models for medical anomaly detection," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2022, pp. 35–45.

[92] M. F. Beg *et al.*, "Computing large deformation metric mappings via geodesic flows of diffeomorphisms," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 139–157, 2005.

[93] B. B. Avants *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.

[94] W. Yin, J.-J. Sonke, and E. Gavves, "Pc-reg: A pyramidal prediction–correction approach for large deformation image registration," *Medical Image Analysis*, vol. 90, p. 102978, 2023.

[95] G. Balakrishnan *et al.*, "Voxelmorph: A learning framework for deformable medical image registration," *CVPR*, 2018.

[96] Y. He *et al.*, "Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching," *IEEE JBHI*, vol. 26, 2021. [Online]. Available: https://doi.org/10.1109/JBHI.2021.3095409

[97] P. Czajkowski and T. Piotrowski, "Registration methods in radiotherapy," *Reports of Practical Oncology and Radiotherapy*, vol. 24, no. 1, pp. 28–34, 2019.