UNIVERSIDADE D
COIMBRA

Catarina Camacho Caldeira

Development and comparative assessment of general and anatomy-specific models for synthetic computed tomography generation

Thesis submitted to the Faculty of Science and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialisation in Clinical Informatics and Bioinformatics, supervised by Dr. Nickolas Papanikolaou, Dr. José Guilherme de Almeida and Dr. João Miguel Castelhano.

September 2025

1 2 9 0 FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE D COIMBRA

Faculty of Sciences and Technology
University of Coimbra

Development and comparative assessment of general and anatomy-specific models for synthetic computed tomography generation

Catarina Camacho Caldeira

Supervisors:
Dr. Nickolas Papanikolaou
Dr. José Guilherme de Almeida
Dr. João Miguel Castelhano

Thesis submitted to the Faculty of Sciences and Technology of the University of Coimbra for the degree of Master in Biomedical Engineering with specialisation in Clinical Informatics and Bioinformatics.

September 2025

Dedication

This is the most meaningful and fulfilling chapter to write. Until now, this Master's thesis has been the most challenging project of my life, but also the most rewarding. Five years ago, when I moved from Madeira Island to mainland Portugal, I had many hopes and dreams, but no idea how things would turn out. Today, I could not be prouder of what I have accomplished. Several people made this journey easier and I am glad to thank them here.

I thank my supervisors, Dr. Nickolas Papanikolaou and Dr. João Miguel Castelhano, for their guidance and support in this project. I also thank my supervisor, Dr. José Almeida, for his orientation and constant support throughout the work, and for always encouraging me to think critically and follow best practices. I am happy for the opportunity I had to learn from you.

I thank the Champalimaud Foundation, particularly the Computational Clinical Imaging Group, for this opportunity.

I thank my mother, Teresa Caldeira, my aunt, Vera Menezes, and my sister, Mariana Caldeira, for always believing in me and for the unwavering support. I could not ask for a better family.

I thank the love of my life, Gonçalo Medeiros, for being by my side every day, reminding me of my strengths and giving me the courage to keep going.

I thank my lifelong friends, Zita Rocha and Rodrigo Clemente, for their kind and supportive words and their true friendship. I also thank Francisca Melim for her companionship during all those long days at the library.

I am grateful to all the friends with whom I have shared these five years in Coimbra. These were remarkable years, thanks to the people I have met and the amazing friends I have made.

Another milestone in my career has been achieved and I am super excited for all that is yet to come.

Resumo

Cerca de 50-60% dos doentes oncológicos são submetidos a radioterapia, cujo planeamento requer a aquisição de imagens de planeamento de tomografia computadorizada (TC) e ressonância magnética (RM). No entanto, esta abordagem multimodal expõe o paciente a mais radiação devido à aquisição adicional da TC, aumenta o desconforto do paciente, tem custos e exigência de recursos adicionais, além de ser necessário um co-registro preciso entre os diferentes tipos de modalidades de imagem. Para ultrapassar estas limitações, estudos recentes têm explorado modelos robustos de aprendizagem automática capazes de gerar imagens semelhantes a TC a partir de RM, abrindo caminho para o planeamento da radioterapia baseado apenas em RM. Diversas arquiteturas têm sido usadas na literatura, incluindo a rede neuronal convolucional em forma de U (U-Net) e variantes das redes geradoras adversariais (GAN), com uma tendência emergente para a integração de diferentes sequências de RM e abordagens híbridas de modelos. Contudo, a geração realista de TC sintéticas continua limitada por conjuntos de dados pequenos e homogéneos, o que compromete a capacidade de generalização dos modelos. Neste trabalho, foi desenvolvido e avaliado um modelo 2D de aprendizagem automática para gerar TC sintéticos realistas a partir de imagens de RM, usando dois conjuntos de dados: SynthRAD2023 e SynthRAD2025. Inicialmente, o conjunto de dados pélvicos do SynthRAD2023 foi utilizado para otimizar os hiperparâmetros do Autoencoder (AE) e da conditional GAN (cGAN), comparando os seus desempenhos. O *ensemble* de modelos AE testado em 18 pares de RM-TC da região pélvica adquiriu resultados superiores (PSNR=28.09 ± 1.51 dB, MAE=71.69 ± 14.28 HU, SSIM=0.84 ± 0.03, MS-SSIM=0.86 ± 0.03). Estes resultados foram confirmados com 50 dados de teste multi-centro de todas as regiões (SynthRAD2025), onde o AE superou a *performance* da cGAN. Posteriormente, o modelo AE foi treinado e avaliado em duas abordagens: dados de múltiplas regiões e dados de uma região específica. Os resultados revelaram melhor desempenho no modelo treinado com dados de múltiplas regiões, sendo a avaliação de teste baseada em métricas de qualidade de imagem, de consistência geométrica e de dose em terapia de protões com modulação de intensidade (IMPT) utilizando 49 pares de RM-TC (15 da região da cabeça e do pescoço (HN), 16 da região abdominal (AB) e 18 da região torácica (TH)). As métricas de imagem foram superiores na região AB (PSNR=25.69 ± 1.99 dB, MAE=106.99 ± 30.93 HU, SSIM=0.75 ± 0.06, MS-SSIM=0.79 ± 0.09), em comparação com as regiões TH e HN. As taxas de aprovação gamma (2%/2mm) foram melhores na região HN, com 98.58 ± 1.12%, o que se compara favoravelmente com a literatura. Embora o modelo demonstre potencial para o planeamento de radioterapia baseado apenas em RM na IMPT, particularmente na região HN (sequências de

RM ponderadas em T1), é necessário trabalho adicional para melhorar a geração de osso nas imagens.

Keywords: Planeamento de radioterapia, Autoencoder, Rede geradora adversarial condicional, Tomografia computadorizada, Ressonância magnética.

Abstract

About 50-60% of cancer patients receive Radiation Therapy (RT), whose planning requires the acquisition of both planning Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) scans. However, this dual-modality approach exposes the patient to more radiation from the extra CT scan, increases patient discomfort, has additional costs and resource demands, and requires accurate co-registration between image types. To overcome these limitations, recent studies have explored robust machine-learning models capable of generating Synthetic Computed Tomography (sCT) images from MRI, paving the way for Magnetic Resonance (MR)-only RT planning. Several architectures have been used in the literature, including U-shaped Convolutional Neural Network (U-Net) and Generative Adversarial Network (GAN) variants, with an emerging trend of integrating different MRI sequences and hybrid model approaches. Nevertheless, realistic sCT generation remains limited by small and homogeneous datasets, which compromises the model's generalisability. In this work, we developed and evaluated a 2D machine-learning model to generate realistic sCT volumes from MRI images using two datasets: SynthRAD2023 and SynthRAD2025. Initially, SynthRAD2023 pelvic data was used to optimise the Hyperparameters (HP) of an Autoencoder (AE) and a Conditional Generative Adversarial Network (cGAN), and compare their performances. The ensemble of AE models tested with 18 pelvic MRI-CT pairs achieved superior results (Peak Signal to Noise Ratio (PSNR)=28.09 $\pm$ 1.51 decibels (dB), Mean Absolute Error (MAE)=71.69 $\pm$ 14.28 Hounsfield Unit (HU), Structural Similarity Index Measure (SSIM)=0.84 $\pm$ 0.03, Multi-Scale Structural Similarity Index Measure (MS-SSIM)=0.86 $\pm$ 0.03). These findings were confirmed with 50 multi-centre test data from all regions (SynthRAD2025), where AE outperformed cGAN. Subsequently, the AE model was trained and evaluated in two approaches: multi-region and region-specific data. Results revealed greater performance in the model trained with multi-region data, where the test evaluation was based on image quality metrics, geometric consistency metrics, and dosimetry in Intensity Modulated Proton Therapy (IMPT) using 49 MRI-CT pairs (15 from the Head and neck (HN) regions, 16 from the Abdominal (AB) region, and 18 from the Thoracic (TH) region). The imaging metrics were superior for the AB region (PSNR=25.69 $\pm$ 1.99 dB, MAE=106.99 $\pm$ 30.93 HU, SSIM=0.75 $\pm$ 0.06, MS-SSIM=0.79 $\pm$ 0.09), when compared to the TH and HN regions. Gamma pass rates (2%/2mm) were better in the HN region, with 98.58 $\pm$ 1.12%, which compares favourably with published work. While the model shows potential for MR-only RT planning workflows on IMPT, particularly for the HN region (T1-weighted MR sequences), further work is needed to improve bone generation.

Keywords: Radiotherapy planning, Autoencoder, Conditional generative adversarial network, Computed tomography, Magnetic resonance imaging

Contents

List of Figures

List of Tables

List of Acronyms

AB Abdominal. vii, xvi, xix, xxi, xxii, 3, 35, 36, 37, 50, 51, 53, 68, 69, 72, 73, 74, 75, 77, 82, 84, 85, 86, 87, 92, 93, 94, 95, 96

ABM Atlas-based Method. 2, 23, 24, 25, 26

AE Autoencoder. vii, xiv, xv, xix, xxi, 3, 18, 19, 20, 44, 46, 47, 49, 50, 58, 59, 61, 65, 66, 67, 68, 69, 71, 72, 73, 74, 75, 76, 82, 91, 92, 93, 95, 96, 97

BDM Bulk-density Method. 2, 23, 24, 26

CBCT Cone Beam Computed Tomography. 33, 35

cGAN Conditional Generative Adversarial Network. vii, xiv, xv, xx, xxi, 3, 19, 20, 21, 26, 27, 28, 29, 32, 44, 45, 46, 47, 49, 50, 58, 59, 61, 62, 63, 64, 66, 67, 68, 69, 70, 71, 72, 91, 92, 93, 95, 97

CNN Convolutional Neural Network. 93

CREPS Content and Style Representation for Enhanced Perceptual Synthesis. 25, 27, 30

CT Computed Tomography. vii, xiv, xvii, xix, xx, xxii, 1, 2, 3, 5, 6, 7, 11, 12, 13, 17, 18, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 39, 40, 41, 43, 44, 46, 48, 50, 51, 52, 53, 54, 55, 57, 79, 82, 83, 84, 87, 89, 91, 92, 93, 94, 95, 96, 97

CTV Clinical Target Volume. 6, 26

CV Cross-validation. 26, 28, 46, 49, 50, 59, 66, 68, 95

dB decibels. vii, xvi, 27, 71, 74, 75, 77, 80, 81, 93

DDPM Diffusion Denoising Probabilistic Models. 27, 29

DL Deep Learning. 23, 26, 32, 43

DLM Deep Learning Method. 2, 23, 25, 26, 28, 32

DVH Dose-volume Histogram. xix, 25, 26, 28, 31, 32, 55, 85, 86, 87, 94, 96

eCNN Efficient Convolutional Neural Network. 26, 28, 29

GAN Generative Adversarial Network. vii, 18, 19, 20, 21, 25, 26, 27, 28, 29, 30, 32, 91, 96

GRE Gradient echo. 15

GTV Gross Tumor Volume. xvi, xvii, 6, 17, 53, 54, 55, 86, 87, 88, 89, 94

HD95 Hausdorff Distance 95th Percentile. xvi, 52, 82, 84, 93

HLUT Hounsfield look-up table. xiii, 13, 14, 53

HN Head and neck. vii, xvi, xvii, xix, xx, xxi, xxii, 3, 34, 35, 36, 37, 50, 51, 52, 53, 54, 68, 69, 70, 72, 73, 74, 75, 76, 77, 78, 79, 82, 84, 85, 86, 88, 92, 93, 94, 95, 96

HP Hyperparameters. vii, xix, 3, 25, 26, 43, 44, 46, 47, 48, 49, 59, 61, 91, 97

HU Hounsfield Unit. vii, xiii, xvi, xxii, 12, 13, 14, 17, 18, 23, 25, 26, 27, 39, 51, 53, 54, 66, 71, 73, 74, 75, 77, 79, 80, 81, 82, 92, 93, 96

Hz Hertz. 14

IMPT Intensity Modulated Proton Therapy. vii, xvii, 9, 53, 89, 93

IMRT Intensity Modulated Radiation Therapy. xiii, 10, 11, 53

KLD Kullback-Leibler Divergence. 25, 26, 30

L1 Least Absolute Deviations. 25, 26, 30

L2 Least Square Errors. 25, 26, 30

LINAC Linear Accelerator. xiii, 7, 8

MAE Mean Absolute Error. vii, xv, xvi, xx, 25, 26, 27, 28, 30, 32, 46, 47, 49, 51, 59, 62, 64, 65, 66, 68, 69, 71, 73, 74, 75, 77, 78, 80, 81, 82, 85, 92, 93, 96

MC-DDPM MRI-to-CT Diffusion Denoising Probabilistic Models. 27, 28, 29, 96

mDICE Multi-class Dice. xvi, 52, 82, 83, 84, 93

ME Mean Error. 25, 26, 28, 31, 32

MeV Mega Electronvolt. 7

MHz Megahertz. 14, 15

MLC Multileaf collimator. xiii, 8, 10, 11

MR Magnetic Resonance. vii, 1, 2, 3, 5, 14, 15, 20, 23, 32, 33, 40, 54, 93, 94, 95, 97

MRI Magnetic Resonance Imaging. vii, xiv, xix, xxi, 1, 2, 3, 5, 6, 7, 11, 14, 15, 16, 17, 18, 20, 21, 23, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 46, 48, 50, 69, 70, 91, 92, 93, 94, 95, 96

MS-SSIM Multi-Scale Structural Similarity Index Measure. vii, xv, xvi, xx, 25, 28, 31, 46, 49, 51, 59, 62, 64, 66, 67, 68, 69, 71, 73, 74, 75, 77, 78, 80, 81, 82, 92, 93

MV Megavoltage. 7

NCC Normalized Cross-correlation. 25, 27, 28, 31

OAR Organs at Risk. xvi, xvii, 1, 2, 6, 7, 9, 11, 17, 53, 54, 55, 86, 87, 88, 94

PBM Patch-based Method. 2, 23, 24, 25, 26, 28

PCC Pearson Correlation Coefficient. 25, 26, 28, 31

pCT Pseudo-computed Tomography. 23, 29, 30, 93

PET Positron Emission Tomography. 6

Pix2Pix Pixel-to-pixel. 26, 27, 28, 29

PL Perceptual Loss. 25, 26, 27, 28, 30, 47

PSNR Peak Signal to Noise Ratio. vii, xv, xvi, xx, 25, 27, 28, 31, 32, 46, 49, 51, 59, 62, 64, 66, 68, 69, 71, 73, 74, 75, 77, 78, 80, 81, 82, 92, 93

PTV Planning Target Volume. 6, 7

ResNet Residual Network. 26, 28, 29, 32

RF Radiofrequency. xiv, 14, 15, 16, 17

RSP Relative stopping power. xiii, 13, 14

RT Radiation Therapy. vii, 1, 2, 5, 6, 7, 9, 10, 11, 13, 17, 23, 25, 26, 27, 32, 33

sCT Synthetic Computed Tomography. vii, xvii, xix, xx, xxi, xxii, 1, 2, 3, 18, 20, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 44, 48, 49, 51, 52, 53, 54, 55, 68, 70, 73, 74, 75, 76, 82, 83, 84, 86, 87, 89, 91, 92, 93, 94, 96, 97

SE Spin-echo. 15

SSIM Structural Similarity Index Measure. vii, xv, xvi, xx, 25, 27, 28, 31, 32, 46, 49, 51, 59, 62, 64, 66, 67, 68, 69, 71, 73, 74, 75, 77, 80, 81, 82, 92, 93

Swin-Vnet Shifted-window Transformer V-net. 27

T Tesla. 14, 35

TE Echo time. xiv, 15, 16

TH Thoracic. vii, xvi, xvii, xix, xxi, xxii, 3, 35, 36, 37, 50, 51, 53, 68, 69, 72, 73, 74, 75, 77, 78, 82, 84, 85, 86, 87, 88, 92, 93, 94, 95, 96

TR Repetition time. xiv, 15, 16

U-Net U-shaped Convolutional Neural Network. vii, 25, 26, 28, 29, 32, 44, 47, 93

V-Net V-shaped Convolutional Neural Network. 26, 28, 29

VMAT Volumetric Modulated Arc Therapy. xiii, 10, 11

Chapter 1

Introduction

This chapter includes a brief explanation of the importance of the topic, context, and the main questions addressed in this thesis. Section 1.1 explains the study motivation, and section 1.2 its context. The main goals are presented in section 1.3, and finally, the organisation of this document is summarised in section 1.4.

1.1 Motivation

According to the World Health Organization, cancer is still a leading cause of death worldwide, with nearly 10 million deaths in 2020, among which 1.80 million were caused by lung cancer and 769.000 by stomach cancer. Despite the high mortality of some cancer types, oral cancer is one of the most common and has a high probability of being cured when detected early and treated properly [1].

From multidisciplinary cancer care, RT is a fundamental and widely used treatment, prescribed to at least one out of two cancer patients. With cancer incidence continuing to rise, a substantial increase in the number of patients needing radiotherapy is expected in the years to come [2].

In addition, planning RT is a crucial step towards an effective dose delivery to the tumour without compromising surrounding healthy tissues. With advanced techniques such as machine-learning methods, only a MR scan (needed for tumour and OARs delineation) is required to generate a sCT (needed for dose calculation), unlike traditional RT planning, which also requires a CT acquisition. This eliminates the need for co-registration between CT and MRI, which implies anatomical and spatial alignment and can introduce errors in dose planning. Thus, exploring a solution based on machine-learning models enables more precise dose calculation and a more efficient workflow by removing an additional imaging modality acquisition (CT), thereby reducing patient burden and demands on medical staff.

When applying machine-learning methods across heterogeneous datasets, potential variations in model performance emerge due to differences in acquisition centres and MRI scanner vendors, a limited number of images per region, and variations in tissue texture, contrast and

anatomical shape. Consequently, more efforts are required to develop and achieve models that are robust to such differences.

1.2 Context

For many cancer locations, the combination of CT with MRI is becoming the current norm for RT planning, due to complementary advantages. In this multimodal approach, CT is used as the main modality as it provides high spatial resolution and electron density information, required for dose computation. MRI is the secondary modality, which has superior soft tissue contrast, facilitating the contouring of lesions and surrounding structures by reducing ambiguities in intra- or inter-observer outlines. However, to simplify the treatment planning process by eliminating the need for multiple scans, there has been a shift toward the adoption of MRI-based planning workflows over conventional CT-based ones [3].

Firstly, using MRI as the primary modality for treatment planning eliminates the need for co-registration (spatial alignment) between CT and MR scans, which is normally required to transfer delineations obtained from MR images onto CT scans for dose calculation purposes, but has been proven to introduce geometric uncertainties that are systematically propagated throughout treatment [3]. By eliminating this source of error, the use of MR images for accurate target and OARs delineation, along with transferring these contours to sCT, which was derived directly from MR image, improves treatment planning accuracy. Furthermore, excluding planning CT scans from the treatment planning process reduces patient exposure to ionising radiation and addresses patient discomfort, while also improving both efficiency and cost-effectiveness, as the patient only needs to undergo a MRI scan.

Thus, ongoing research efforts have focused on generating sCT scans from MRI to enable MR-only treatment planning. Methods of this image-to-image translation can be divided into four groups: Bulk-density Method (BDM), Atlas-based Method (ABM), Patch-based Method (PBM) and Deep Learning Method (DLM). In the BDM, MRI is segmented into different classes, assigning each a density value derived from CT data, but requiring manual segmentation of structures [4][5]. ABM aligns the target MRI to previously known CT-MRI atlas pairs through deformable registration to generate sCT images, but its performance is heavily dependent on accurate co-registration [4][5][6][7]. The PBM is a machine-learning approach similar to the ABM with four additional steps and a generation time of minutes. However, it is influenced by imprecise inter-patient registration [4][5][8]. At last, the most recent and growing area of research involves using DLM to solve this image synthesis task, due to rapid generation of realist sCT images without requiring multimodal registrations [4][5][9].

Although machine-learning methods are a promising approach, it is acknowledged that refinement of sCT methods is still needed due to treatment site dependence, limitations of current techniques and the need for further validation [3].

1.3 Research goals

This work addresses the gap in the literature regarding training and evaluation of machine-learning models with limited dataset sizes and a lack of multi-regional data. By using a larger and more diverse dataset from the AB, TH and HN regions, several models were trained on these individual regions and on combined multi-region data to evaluate the superiority of one approach in comparison to the other.

The goal of the present work is to investigate efficient models using the AE and cGAN to generate sCT images from MRI. In this thesis, 18 pelvic MRI-CT pairs from the SynthRAD2023 test set are used to evaluate the model trained on the SynthRAD2023 development set (after HP search). Subsequently, 49 MRI-CT test pairs from the SynthRAD2025 dataset - 15 from the HN region, 16 from the AB region and 18 from the TH region - are used for a second validation to draw conclusions about general and region-specific approaches. The main goals of this work are the following:

- To explore the use of general and region-specific 2D models for the MRI-to-CT synthesis.
- To generate sCT volumes suitable for treatment planning, from head and neck, thoracic and abdominal regions.
- To evaluate the quality of the synthetic-CT in three domains: image quality, geometric consistency, and dose calculation using proton therapy plans.

1.4 Structure

This document is organised into five chapters beyond the introduction.

Chapter 2 presents the background concepts of radiotherapy workflows in general and in MR-only, imaging modalities in radiotherapy, and deep learning.

Chapter 3 reviews the state of the art related to the different types of sCT generation, including deep learning models.

Chapter 4 describes all the methods investigated. In addition, it contains the description of the dataset, how it was processed, the architectures employed and the training protocol, as well as the performance assessment.

Chapter 5 reports a detailed description of the results regarding image quality metrics, geometric consistency metrics, and dose metrics.

Chapter 6 discusses the results thoroughly and addresses potential future research in the field.

Chapter 7 presents a conclusion of the thesis work.

Chapter 2

Background Concepts

This chapter introduces the background concepts for a better understanding of the study and is divided in four main sections. Firstly, section 2.1, describes the entire radiation therapy planning workflow. Section 2.2, presents a brief overview of the fundamentals of both CT and MRI, and the comparison between the two. Section 2.3 describes the workflow of the MR-only radiotherapy planning. Lastly, section 2.4 reports the concepts of deep learning and architectures used in this work.

2.1 Radiotherapy planning workflow

2.1.1 Treatment planning

Once the patient has a confirmed diagnosis, the malignant tumour can be locally treated with radiation, which is a complementary treatment to chemotherapy and hormone therapies [10]. RT uses ionising radiation, composed of ions (electrically charged particles) to deposit energy in cells, either killing cancer cells directly or damaging their genetic material, resulting in cell death [11]. The aim of RT is to maximise the radiation dose to the target tumour and minimise exposure to normal cells that are adjacent to cancer cells or in the radiation path. This process involves selecting a reproducible patient positioning, localising and delineating the tumour and surrounding structures, choosing the irradiation geometry, calculating the dose for the required plan optimised through objective functions and dose constraints specific to the structures, and finally evaluating the resulting dose distribution [10].

As the effects of high-energy radiation from RT are irreversible, planning is vital. Initially, a simulation appointment is scheduled to simulate the patient position (usually supine) for RT so it is reproducible for each day of treatment. This approach ensures both the effectiveness and security of treatment for individual radiation oncology patients by reducing radiation exposure to normal tissues. During simulation, various immobilisation devices are utilised depending on the treated region to guarantee a consistent patient position during planning and treatment, while minimising intra-fraction motion. For instance, for the head and neck region, the thermoplastic mask is the most commonly used immobilisation device [12], while in the thoracic region (e.g.,

lung cancers), the immobilisation systems include polyurethane foam casts or evacuated vacuum cushions placed beneath the patient's thorax, combined with a positioning device to hold the patient's arms overhead [13]. Concurrently, the overlay of the lasers on the patient's body is marked. After this, patients undergo CT imaging, which is the backbone of 3D RT, used to image the tumour and adjacent anatomy, facilitating the design of 3D precise treatment by the doctors [14]. However, inadequate soft tissue contrast on the CT scans led to the incorporation of other imaging modalities, such as MRI and Positron Emission Tomography (PET) scans, into the definition of target tumour. Nowadays, MRI-simulators designed for treatment planning are commercially available as they provide a better option than diagnostic scanners [15]. Unlike diagnostic MRI scanners, MRI-simulators include larger tunnel sizes, flat tabletops designed to accommodate immobilisation devices, external laser systems, and specific imaging protocols.

High-quality imaging scans, such as CT and MRI, are crucial for defining standardised concepts. In RT planning, three essential target volumes are considered: GTV, Clinical Target Volume (CTV) and Planning Target Volume (PTV) as depicted in Figure 2.1, alongside OARs. The gross target tumour, which represents the identified tumour mass, is delineated and termed GTV. In order to take into account the future microscopic spread of the tumour, further margins are included, resulting in the CTV. Moreover, a supplementary volume expansion is reported as PTV to account for the daily fluctuations in tumour motion and patient positioning, as well as the uncertainties in planning. The treated volume comprises the region intended to receive the radiation dose, covering the PTV with as little excess as possible. Thus, OAR are exposed to minimal irradiation as they include normal tissues and relevant structures near the tumour [14]. After the delineation of the treated volume, the radiation dose is calculated. At the end, the beam size, number, trajectory, and weighting (measures the correlation between the absorbed dose and clinical effects) towards these structures is optimised [10]. Once the radiotherapy planning volumes are established, the physician collaborates with dosimetrists and medical physicists to develop an optimal radiation delivery plan, which is then prescribed to the patient [16] [17].

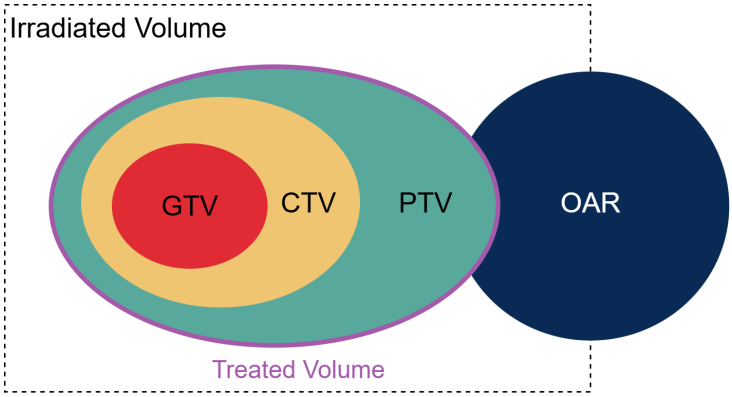


Figure 2.1: Schematic representation of a treated and irradiated volume. Abbreviations: GTV - gross tumour volume, CTV - clinical target volume, PTV - planning target volume, OAR - organ at risk.

Combining planning CT intensity values and electron density information, as well as the geometric integrity of CT images with the superior soft tissue contrast of MRI, offers significant advantages, but involves an accurate co-registration of planning CT-MRI. In practice, after uploading the images into the treatment planning system, the MRI obtained from the simulation appointment is registered to the patient's CT simulation image to facilitate contouring of the PTV and OARs volumes [10].

While manual co-registration is time-consuming and requires expertise along with consistent practice, automated registration is more efficient, but still demands a highly skilled professional to detect and correct potential errors. Image co-registration consists of aligning the images from different modalities into one coordinated system. A clear example of this in RT is the transformation of the MRI images into the coordinated system of the CT images, in order to match the anatomy of the CT data [15]. Various factors, including motion and anatomical changes between time-separated acquisitions, determine the complexity of the multimodal image registration. Although the majority of RT clinics utilise only rigid registration at treatment planning and delivery, the registrations can be rigid, affine, or deformable. Rigid registration transforms the images by translation and rotations in all directions. Affine registration includes the transformations of the rigid registration, but with the addition of scaling, shearing, and plane reflection. In contrast, deformable registration applies localised spatial transformations, which means that different parts of the image move independently to improve the alignment between images. For example, deformable registration is applied to pelvic or abdominal images, where internal organ motion is expected [18].

2.1.2 External beam radiation delivery

The most common approach to deliver radiation in a clinical setting is external beam radiation. Through this method, the radiation is delivered from outside the body to the tumour with high-energy rays (photons, particle radiation or protons) [11]. Linear Accelerator (LINAC) machines are a modern type of external beam radiation therapy that delivers precise and controlled doses of radiation to target tumours while minimising damage to nearby healthy tissues. The radiation delivery process in LINAC (Figure 2.2 and Figure 2.3) starts in the electron gun, where a heated filament (cathode) emits a cloud of electrons. These electrons are then accelerated by an electric field between the cathode and a thin metal window (anode), in the accelerator waveguide, reaching energies of up to 18 Mega Electronvolt (MeV). As soon as the electrons reach the bending magnet, their path is redirected toward the centre of the beam's gantry axis of rotation (or patient), known as the isocentre. Next, these high-energy electrons from the target either collide with a heavy metal target to produce X-rays of > 6 Megavoltage (MV) for **photon treatments** (X-ray mode) or strike a scattering foil for **electron treatments** to spatially distribute the electron beam (electron mode) [19] [20]. In X-ray mode, the X-rays produced at the target are concentrated in a bullet-shaped beam form by a fixed primary collimator and then pass through the cone-shaped flattening filter to create a uniform beam. In contrast, for electron treatments, the target and flattening filter are replaced by an electron scattering foil, all housed in a circular mechanism, the carousel, located below the target. Furthermore, the

role of the ionising chambers is to monitor beam flatness and symmetry, as well as dose output, expressed in monitor units (MU), ensuring reproducible delivery of the prescribed dose during treatment. After passing through the ion chamber, the radiation beam is further shaped by a secondary dynamically movable collimator that includes two pairs of jaws and the MLC. By rotating the gantry around the patient and moving the treatment couch on which the patient lies, radiation is precisely delivered to the tumour from multiple angles [20][21].

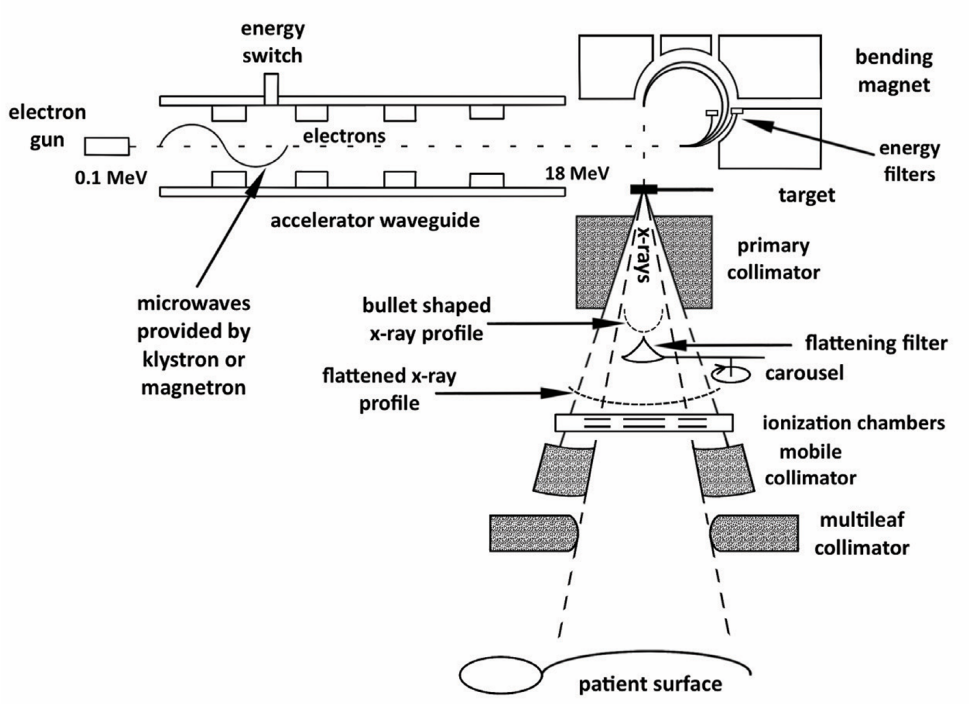


Figure 2.2: Schematic of LINAC delivery systems. Adapted from [20].

Figure 2.3: Medical LINAC machine. Adapted from [20].

Proton beam therapy is another form of external radiotherapy that uses positively charged particles (protons) with a well-defined range of penetration into the tissue. Once protons

are accelerated, they are guided to the gantry to irradiate the patient. Upon producing the desired beam energy, this is spread out as a mono-energetic beam, which deposits its energy at a specified depth. In other words, most of proton's energy is deposited at the end of its path, resulting in a rapid increase in proton dose at a specific region known as the Bragg Peak (Figure 2.4). While a mono-energetic beam covers a small part of the tumour, proton beams with multiple energies can be combined using a modulator to spread out Bragg peaks across multiple depths. A major advantage of the proton radiotherapy is the tumour-centred beam irradiation, which minimises exposure to OARs. As a result, proton RT reduces toxicity to healthy tissues, unlike X-ray therapy.

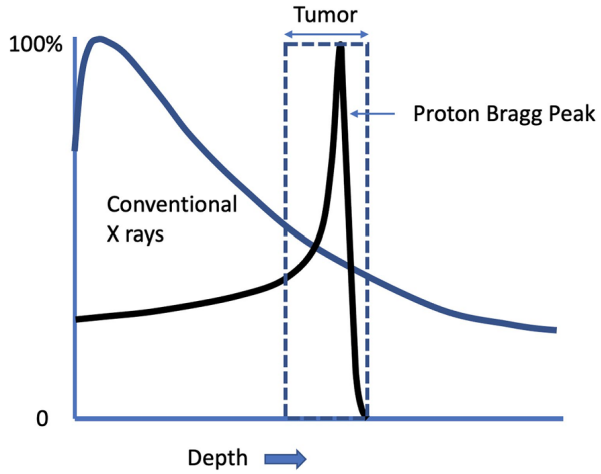


Figure 2.4: Bragg peak diagram. When accelerated, protons acquire kinetic energy that is deposited within the tumour, which increases the percentage of energy represented by the y-axis, contrary to conventional X-rays. Then, their energy is quickly lost through interactions with atoms, stopping the delivery of radiation. Based on an original image from [22].

2.1.3 Treatment planning strategies

To achieve more conformal dose distributions in the tumour, proton beam scanning can be combined with intensity modulation, a technique known as IMPT [23]. IMPT, also referred to as the "pencil beam" or "active scanning", is a more sophisticated and complex method of delivering RT. These narrow proton beams, resembling "pencils", originate in the accelerator and are manipulated to reach the tumour in layers of spots at different depths. The beam is modulated by changing three parameters: the number of protons, which determines the local dose deposition, the energy in local penetration, and the magnetic deflection. By using magnetic fields to steer the charged particles in different directions away from the central axis, the fields enable IMPT to reach various spots and cover the entire surface of the tumour [24]. Since the IMPT relies on electromagnetic control of the pencil beam to cover the tumour, while minimising radiation exposure to surrounding healthy tissues, it offers a promising treatment for cases where dose escalation is needed alongside sparing of the OARs. However, a significant challenge of this technique is managing organ motion in regions such as the abdomen and thorax. Because IMPT delivers proton beams to multiple layers of the tumour, even small movements can compromise treatment accuracy. Therefore, IMPT is a much preferred option for treating tumours in areas

with minimal motion, such as the head and neck, spinal cord, lower pelvis, and certain lung cancers with a movement less than 5 mm [23].

Intensity Modulated Radiation Therapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT) are two types of radiation therapy techniques designed to optimise the cumulative dose distribution delivered to the patient from different directions, using non-uniform radiation fluence. This optimal fluence for a given set of beam directions is computed through a process called inverse planning, which requires a treatment planner to create the optimisation criteria [25]. In forward treatment planning, firstly, the geometry of the beam, such as its orientation, shape, modifier, and weights, is defined, and then the 3D dose distribution is calculated, repeating the process until the resulting dose is satisfactory. Conversely, inverse planning begins with the definition of a desired dose distribution, subjected to optimisation criteria (objective function and constraints), and then the planning parameters are calculated [26]. The process is repeated until a feasible solution is found. To comply with the predefined dose distribution criteria, inverse planning is performed using specialised software that divides each radiation beam into beamlets (subdivisions of a radiation beam) and adjusts their weights accordingly. The resulting optimally modulated fluences are translated into MLC leaf sequences, and then electronically transmitted to a linear accelerator that delivers the calculated intensity-modulated beams.

MLC consists of several leaf pairs, where each pair has two opposing leaves, forming a sliding window. As the leaf pairs move independently (at different speeds), the width of the window changes, exposing different points of the patient's body to radiation in different time periods. Thus, these points are irradiated with different doses. This represents a fundamental principle of how intensity modulation is achieved during the RT treatment delivery. The MLC scheme and principle of intensity modulation are shown in Figure 2.5.

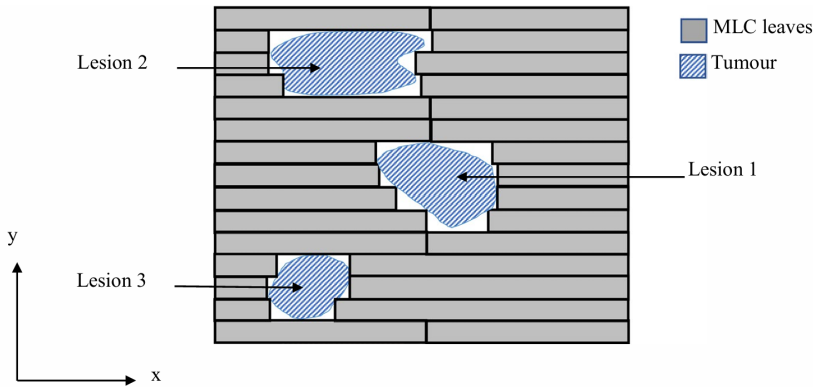


Figure 2.5: Schematic representation of delivering IMRT/VMAT using MLC. Figure adapted from [27].

The main difference between IMRT and VMAT lies in the delivery method: VMAT delivers radiation continuously as the linear accelerator gantry rotates around the patient, whereas in IMRT, the radiation delivery is static, with the linear accelerator rotating between beam deliveries [14]. In more detail, during VMAT, the radiation beam is always active, and the MLC

continuously moves to adjust its shape to each subfield (individual segments within each arc) as the gantry rotates. Per gantry angle, an arc is defined to deliver one subfield. The overlap of subfields, through multiple arcs, results in the modulation of radiation therapy at each beam angle. IMRT delivery is performed through the "step-and-shoot" (or "stop-and-shoot"), where radiation is delivered in multiple static subfields shaped by the MLC, with the gantry stopped and the beam turned off while the MLC leaves reposition to form the next subfield. Another delivery method is the sliding window technique, where MLC leaves move unidirectionally while the beam remains on [25]. The pictures in Figure 2.6 illustrate examples of IMRT and VMAT plans.

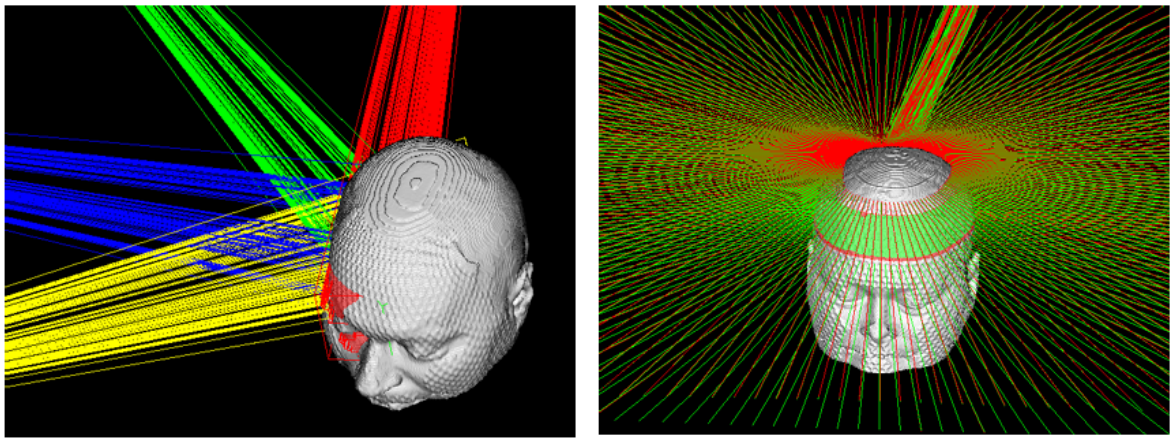


Figure 2.6: Examples of beams arrangement in IMRT (left), and full arcs in VMAT (right). Figure adapted from [28].

2.2 Imaging modalities in radiotherapy

Imaging plays an essential role in cancer care, from allowing clinicians to discover the tumour and determine its stage by evaluating certain anatomical patterns within the scans, to guiding radiotherapy planning. For RT planning, the patient must undergo appropriate imaging modalities, in which the full anatomical location and physiological characteristics of the tumour, and surrounding OAR are assessed. After treatment, imaging remains crucial to monitor potential recurrences and continued surveillance of the cancer [29].

In recent years, advances in technology have led to a sophisticated range of imaging modalities that provide detailed information on tumour morphology, anatomy, biology, and patterns of cancer spread with greater precision. Among these, the widespread availability of CT and MRI in patient care has been essential to the practice of RT, including its planning. However, CT imaging exposes patients to higher radiation doses, so it is important to ensure that these are used appropriately and optimally [29].

2.2.1 Computed tomography

During scanning with CT, the X-ray tube emits beams that pass through axial sections of the patient's body from different angles as it rotates around the patient. The detectors placed around the patient capture the X-rays that exit the body and measure their intensity by converting them into electric signals. These signals are then amplified and processed to account for variations in the detector system, such as misalignments between the detector and the X-ray tube. The processing can also include corrections for beam hardening artefacts, which are caused by a higher absorption of low-energy photons compared to high-energy photons when X-rays pass through high-density objects, such as metal [30]. After this, the data from the detector is transformed into X-ray attenuation values based on the original intensity of each ray, resulting in CT raw data (in 2D). Raw data projections are then mathematically filtered and reconstructed into a cross-sectional image, typically using the filtered back projection algorithm [29]. In this algorithm, the raw data is first filtered to reverse image blurring. This is achieved by convolving the raw data with a convolution kernel, which produces filtered projections. Subsequently, these are backprojected onto the image matrix to generate the reconstructed CT image. The resulting image matrix is normalised and converted to integer values to meet the specifications of the display hardware, before storing and displaying the reconstructed CT image [31]. Since the reconstruction algorithm's main goal is to calculate the attenuation coefficient for each pixel in the image, this coefficient is then converted into a CT number (HU), which is included in the final clinical CT image, defined as:

$$\mathrm{CT} = 1000 \times (\mu_{\mathrm{material}} - \mu_{\mathrm{water}})/\mu_{\mathrm{water}} \tag{2.1}$$

where μ_{material} and μ_{water} represent the CT linear attenuation coefficients of the material or water, respectively. [29].

Based on the previous equation, several CT numbers are defined in a scale, for example, -1000 HU corresponds to air, 0 HU represents water, and around 50 HU represents soft tissue (without an upper limit). While the acquisition of CT produces cross-sectional 2D images of the body, they can be transformed into 3D gray images through axial stacking [29]. CT scanners and treatment planning systems typically store data at 12-bit depth, which means $2^{12} = 4096$ values, ranging from -1024 to 3071 HU, and display about 8 bits, which is 256 shades of gray [32].

In clinical practice, CT is the current conventional 3D imaging modality for radiation oncology simulation. The choice of CT scans for the planning of radiation therapy focuses on two key reasons: (1) inherently electron density information, indispensable for dose calculation; (2) geometrically robust, enabling accurate targeting of tumours during treatment planning and delivery [14].

CT values to relative stopping power calibration

As CT scans provide information about how different tissues attenuate X-rays, their varying densities can be determined. This enables treatment planning systems to accurately account for

tissue heterogeneities and ensure accurate dose delivery. Since HU numbers derived from CT scans for a given tissue depend on the X-ray beam quality and individual scanner parameters, the same tissue can have different HU values. To account for this variability, a calibration process is required between the HU values and an intrinsic property, which differs between photon and proton RT. In photon RT, HU values are mapped to the electron density of a tissue, while in proton RT the calibration is between HU values and RSP, through a HLUT or other calibration methods [33] [34]. The conversion curves are usually determined using phantoms composed of tissue-equivalent materials, simulating the properties of human tissue. Despite differences in calibration targets between proton and photon RT, CT scans in both are acquired using X-rays, and the development of the tissue calibration curve follows a similar procedure. The electron density of a material, which is used in both modalities' calibration, is given by:

$$\rho_\epsilon = \frac{\rho N_g}{\rho_{\text{water}} N_g^{\text{water}}} \tag{2.2}$$

where ρ is the density and N_g is the number of electrons per unit volume of a material composed of chemical elements, i, defined as:

$$N_g = \sum N_g^i = N_A \sum \frac{\omega_i Z_i}{A_i} \tag{2.3}$$

where N_A is Avogadro's number (6.023×10^{23}), Z_i and A_i are the atomic number and atomic weight of the i-th element and ω_i is the weight proportion.

For proton treatment planning, the proton's RSP, ρ_s, that defines the rate at which a proton loses energy in different tissues compared to water, is required for the proton calibration curve and is expressed in the Bethe-Bloch formula:

$$\rho_s = \rho_e \cdot \frac{\log\left(\frac{2m_e c^2 \beta^2}{I_m(1-\beta^2)}\right) - \beta^2}{\log\left(\frac{2m_e c^2 \beta^2}{I_{\text{water}}(1-\beta^2)}\right) - \beta^2} = \rho_e \cdot K \tag{2.4}$$

where $c^2\beta^2$ represents the squared velocity of the proton, c is the speed of light in vacuum, β is the ratio of the particle's velocity to c, m_e is the mass of the electron, I_m and I_{water} are the mean ionisation energy of the material (target atoms) and water.

A clear example of the HU-to-RSP conversion is depicted in Figure 2.7, which represents the default lookup table used by matRad software, an open source treatment planning system for RT [33].

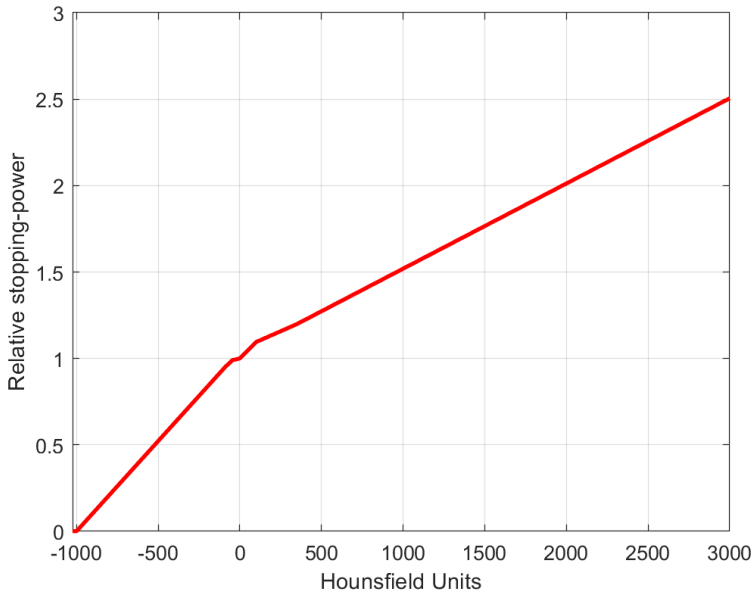


Figure 2.7: HLUT. Graphical representation of the conversion of HU into RSP from matRad.

2.2.2 Magnetic resonance imaging

MRI is based on measuring RF radiation, which is produced as a result of induced transitions between the nuclear spin states of hydrogen atoms when exposed to a strong external magnetic field (Figure 2.8). In this context, the angular momentum of atomic nuclei is commonly referred to as spin. At greater length, the positively-charged protons in the body's water molecules align either in parallel (low energy) or anti-parallel (high energy) with the magnetic field to which they are exposed, depending on their energy state. This strong magnetic field originates from a large magnet within the MR scanner, where superconducting electromagnets are the most commonly used, producing a field strength of 1.5 Tesla (T). Most protons align with the magnetic field (low energy), which is along the main axis (Z) of the scanner, corresponding to the length of the patient, because in this orientation the spins reside in a more favourable energy state. The sum of vectors of the magnetic moments produced by the excess of protons aligned with the magnetic field is termed net magnetisation, M_0. Additionally, the protons that line up with the external magnetic field also rotate or spin (precess) around the Z axis at a frequency that is directly proportional to the field strength, called the Larmor frequency. The Larmor equation represents what has been mentioned:

$$\omega_0 = \gamma B_0 \tag{2.5}$$

where ω_0 represents the angular frequency of rotation, in Hertz (Hz) or Megahertz (MHz), γ is the gyromagnetic ratio (constant specific to the atomic nucleus), and B_0 the external magnetic field strength in T. For protons, the gyromagnetic ratio is $\gamma = 42.58$ MHz/T, which quantifies the relationship between the magnetic dipole moment and the angular momentum. In other words, for each T (strength of magnetic field), the proton's magnetic moment precesses at a

frequency of 42.58 MHz. ω_0 and γ are vector quantities with direction and magnitude [29].

To acquire an image, a RF energy pulse, which is produced with a RF power amplifier, is emitted into the body with a range of frequencies centred around the Larmor frequency. This creates a circularly polarised magnetic field, termed B_1 field, which locally overcomes B_0. Upon this, the spins in the body absorb energy and start rotating in a new direction orthogonal to the net magnetisation vector, around the B_1 field direction. Depending on the energy received by the RF pulses, the protons become excited and their spins can flip orthogonally to their initial position (90^{o} pulse), antiparallel to it (180^{o} pulse), or any angle in between [29].

Finally, when the RF pulse is turned off, the excited spins return to their equilibrium state aligned with the magnetic field, via processes characterised by relaxation times, known as T1 (longitudinal relaxation) and T2 (transverse relaxation). T1 relaxation time is the time that it takes for protons to return to their longitudinal magnetisation (M_0) along B_0 by releasing energy to the surrounding environment. T2 relaxation time is the time that protons take to lose the phase among spins precessing perpendicular to B_0, causing the exponential decay of the MRI signal in the transverse plane. The MRI scanners detect this decaying signal at a rate dependent on the tissue characteristics and process the emitted MR signal to acquire high-resolution images through a mathematical transformation called the Fourier transform [29].

For the MR image acquisition, specific pulse sequences, consisting of RF pulses and gradient waveforms, are chosen to be applied at different times in a particular way to generate images with the desired contrast. There are two main types of MR pulse sequences: Spin-echo (SE) and Gradient echo (GRE). All other sequences are variations of these two. MR pulse sequences can be in 2D, with one section (or slice) acquired at a time, or 3D, with multiple sections obtained in a single acquisition, producing an entire volume. Regarding tissue contrast on images, there are two key parameters: the Repetition time (TR), which refers to the time between the application of successive RF excitation pulses, and the Echo time (TE), which is the time between the application of a RF pulse and the peak of the detected echo (MR signal) [35]. Not only can these two variables be adjusted by the scanner operator, but also the flip angle, which is the angle the M_0 is rotated away from the B_0 by the RF pulse. Changes in these three influence how much the spins relax before signal readout. Once the intrinsic relaxation times, T1 and T2, of different tissues are known, operators can change the image contrast between imaging sequences (SE and GRE), and therefore, assess different physiological characteristics [29].

The most standard images are termed T1-weighted or T2-weighted. Sequences with short TR and short TE are used to obtain T1-weighted images, while those with long TR and long TE result in T2-weighted images [35]. In T1-weighted images, water-based tissues have long T1 and appear relatively dark, while fat tissues have short T1 and appear bright. Gadolinium, a commonly used contrast agent, shortens T1 relaxation times, thereby highlighting tissues that absorb the contrast. In T2-weighted scans, tissues with longer T2 are brighter or hyperintense, which makes it easier to see fluid around tumours, whereas tissues with shorter T2 such as tumours in the prostate peripheral zone are darker or hypointense. Common to both T1-weighted and T2-weighted images, air and cortical bones show low signal intensity, appearing

darker than surrounding regions [29].

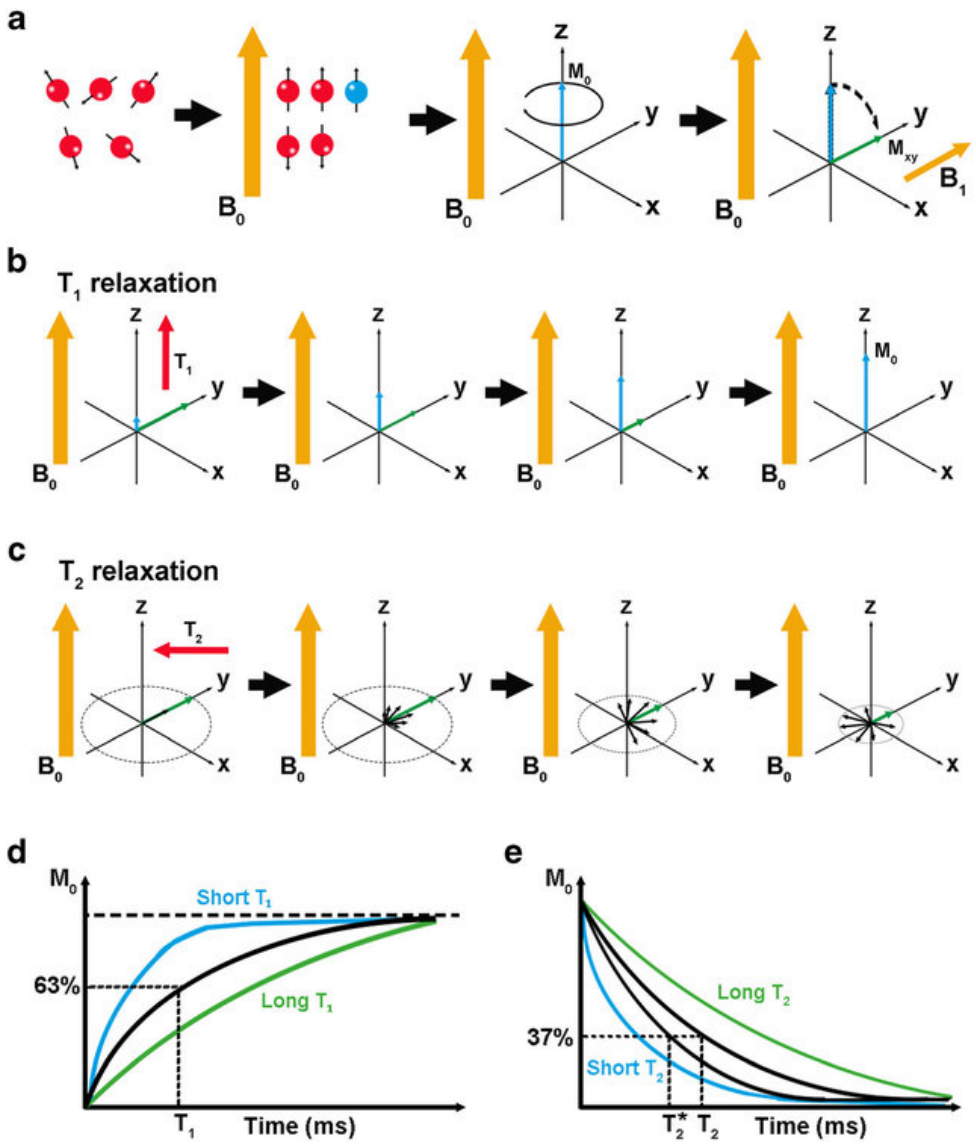


Figure 2.8: Schematic representation of the basic principles of MRI. a) Protons are represented as red balls spinning around their own axis. When an external magnetic field B_0 (orange arrow) is applied, protons tend to align with the direction of B_0 and have only two possible orientations, spin-up and spin-down. The difference between the protons aligned parallel and antiparallel to B_0 represents the protons that are responsible for the MRI signal (blue ball). The sum of these protons can be described by a magnetization vector (M_0, blue arrow). If a second magnetic field (B_1) orthogonal to B_0 is applied, it is possible to tilt M_0 of 90° along the x-y direction (M_{xy}, green arrow). When the RF pulse that originates B_1 is switched off, M_{xy} returns to the equilibrium through two processes: b) T1 and c) T2 relaxation. d) T1 relaxation is defined as the time needed to achieve the 63% of the original longitudinal magnetisation. The blue curve (fat) and the green curve (water) represent tissues with short and long TR (relates to T1) values, respectively. e) T2 relaxation is defined as the time to dephase up to 37% of the original value. Blue curve (fat) and the green curve (water) represent tissues with short and long TE (relates to T2) values, respectively. [36].

2.2.3 Comparison between CT and MR imaging

Currently, the CT-based treatment planning workflow incorporates MRI, which is used to define OAR and the target, since it provides superior soft tissue contrast, making structures easily distinguishable. Therefore, subjectivity is minimised, which significantly reduces inter- and intra-observer contouring variability across different cancerous regions. Precise MRI delineation improves dosimetry and increases the therapeutic ratio (the GTV receives the necessary radiation dose while sparing the surrounding OAR). Furthermore, MRI provides valuable functional information, which can be considered during treatment planning. Regarding CT advantages, image intensity reflects the electron density of different tissues in HU. This is important because the current input variable accepted by treatment planning systems to calculate dose distribution is HU values [37]. As detailed in Table 2.1, the main disadvantage of using MRI in RT is that it cannot provide electron density information, unlike CT. This information is indispensable for dose planning, since it correlates with the way tissues attenuate the radiation beams.

Table 2.1: General comparisons between MRI and CT imaging. Relative scale for comparisons: Nil<Minimal< Present, and Poor< Moderate< Good< Excellent. Abbreviations: Gd - gadolinium, DTPA - diethylenetriaminepentaacetic acid. Table adapted from [10].

Subject	**Parameters**	**MRI**	**CT**
Patient	Magnetic safety concerns	Present	Nil
	RF heat deposition	Present	Nil
	Ionizing radiation dose	Nil	Present
	Claustrophobia in scan tube	Present	Minimal
	Scanning noise	Moderate	Minimal
	Contrast materials allergy:		
	- Iodinated contrast	Not applicable	Present
	- Gd DTPA	Minimal	Not applicable
Characteristics	Soft tissue contrast	Excellent	Moderate
	Cortical bone contrast	Poor	Excellent
	Detection of calcifications	Poor	Excellent
	Metallic artefacts:		
	- Non-ferromagnetic material	Minimal	Present
	- Ferromagnetic material	Present	Present
Machine	Size of tunnel aperture	Smaller	Larger
	Image resolution	Good	Excellent
	Scanning time	Moderate	Short
	Electron density information	Nil	Present
	Functionality and technical sequences	Large	Limited
	Geometrical image accuracy	Distortion present	Excellent
	Multiplanar imaging	Any plane	Limited
	Multiplanar reconstructions	Available	Available
	Cost	Higher	Lower
	Availability	Moderate	Widely available

2.3 MR-only radiotherapy planning workflow

Since the boundaries of structures such as the prostate or brain tumours cannot be identified on CT, the MRI contours are transferred to the CT through image registration. While this dual-imaging approach has its own advantages, it introduces several challenges, concerning the registration of MRI to CT:

- Co-registration complexities: The alignment of MRI with CT introduces geometrical uncertainties (2mm for the brain and 2-3mm for prostate and gynaecological patients) that are systematic, compromising the radiotherapy treatment. Thus, it is crucial to ensure accurate spatial representation to detect lesions and calculate organ-specific dose delivery. However, this process requires expert input and is complicated by factors such as differences in patient positioning between modalities or gas pockets.

- Patient discomfort: The patient is required to do two imaging sessions, exposing them to additional radiation (from CT), discomfort, and a longer time in the hospital.

- Resource and staff demanding: Acquiring multiple imaging modalities requires greater resources and additional medical staff.

Such challenges can be overcome by generating sCT from MRI using robust machine-learning models, speeding up the workflow by demanding only MRI acquisition and reducing the need for registration. Personalised radiotherapy planning can be achieved through sCT data, as it provides the only available HU values indispensable to dosimetry. Furthermore, with those, it is possible to simulate how the individual patient may respond to radiation therapy using computational models. Such simulations can predict tumour shrinkage, potential adverse effects, and optimal dose distribution [37].

2.4 Deep learning

By using multiple interconnected layers, deep learning models can extract features from raw data, process large volumes of information, and learn different relevant aspects to produce meaningful outputs. As a subfield of machine learning, deep learning was structured to mirror the function of the human brain's neural networks using mathematical operations.

The training process consists of adjusting the weights and biases of a neural network by minimising iteratively a loss function that quantifies the discrepancy between the model's predictions and the ground truth. All in all, learning to identify patterns and relationships in the data enables the network to generalise to unseen data. There are several types of deep learning algorithms, each designed for different objectives. For instance, GANs are used for generative modelling tasks while AEs are often used for dimensionality reduction; however, both can be applied to image synthesis [38].

2.4.1 Generative adversarial network and Autoencoder

GANs, are generative models that, in simple terms, take a random noise vector z as input, into their architecture (Generator G), and learn to map it into an output y [39].

$$G : z \rightarrow y$$

This class of generative models consists of two neural networks: a discriminator D and a generator G, which are trained simultaneously, using different optimisation routines to tackle complex generative tasks by initially approaching them as simpler classification problems [40] [41]. At the testing stage, only the trained generator model is used to generate synthetic images. This framework is a game between two players with different objectives:

1. **Generator's Goal**: Starting from random noise, produce samples that are as realistic as possible.
2. **Discriminator's Goal**: Examines the output from the generator to determine whether the samples are real or fake.

During adversarial training, the G and D compete to reach the Nash equilibrium of a two-player game, a state where neither G nor D can improve their performance without altering the other's performance. This occurs when:

- The generator produces samples indistinguishable from real data for the discriminator (i.e., random guessing - a 50% probability of being classified as real by the discriminator applies equally to both real and fake inputs).
- The discriminator cannot improve its performance further, as the generated data matches "perfectly" the real data [42].

This weakens the feedback from the discriminator, hindering GAN convergence if training continues, as the generator learns from increasingly meaningless feedback. As a result, GAN training often encompasses brief and unstable convergence.

A GAN can be interpreted as an AE with an adversarial objective. In this analogy, AE is similar to GAN's generator, as it is a neural network that is trained to reconstruct its input. The AE consists of three main components: an encoder, where the input data is compressed into a lower-dimensional representation; a bottleneck layer, where the essential features of an input are extracted in a representation smaller than the input; and a decoder, where this compressed representation is reconstructed into the original input, aiming to minimise the reconstruction error [43].

2.4.2 Conditional generative adversarial network

The fundamental principle behind cGAN is similar to GAN. However, cGANs learn the mapping to achieve y, from observing the image x and a random noise vector z [39].

$$G : \{(x,z)\} \rightarrow y$$

cGAN constitutes a more complex architecture to generate synthetic-CT images when compared to the AE, due to its adversarial training occurring between two neural networks: the discriminator and the generator.

The cGAN objective, or the min-max loss, is typically defined as:

$$\mathcal{L}_{cGAN}(G,D) = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1 - D(x,G(x,z)))], \tag{2.6}$$

The first term, $\log D(x,y)$, represents the discriminator's correct classifications of real data as real. The discriminator aims to maximise this term in order to increase its accuracy in detecting real data. The second term, $\log(1 - D(x,G(x,z)))$, reflects the discriminator's ability to correctly classify generated images as fake. Since the generator's objective is to fool the discriminator by penalising the discriminator for classifying generated images ($G(x,z)$) as fake, the generator aims to maximise $D(x,G(x,z))$, so that the discriminator is fooled into classifying the produced images as real.

Generator

The cGAN's generator uses feedback from the discriminator to improve its ability to generate synthetic data. Essentially, it adapts to produce outputs that the discriminator is more likely to classify as real. The generator's loss penalises it for failing to produce realistic data, encouraging it to enhance the quality of its outputs. Concerning backpropagation, it flows through both networks but updates only the generator's weights, ensuring the generator learns how its outputs influence the discriminator's response [40].

Discriminator

The discriminator's primary role is to classify the samples generated by the generator as real or fake. During the discriminator's training phase, the generator's weights remain constant, producing only synthetic images for the discriminator's classification. If the discriminator misclassifies a real instance as fake or a fake instance as real, its loss is penalised. The discriminator then updates its weights through backpropagation based on the calculated loss [40].

To enable the model's focus on localised regions, particularly enhancing its effectiveness to detect small and subtle abnormalities, a patch-based discriminator can be utilised. The evaluation of the images is performed on a patch level, sliding a receptive field over the generated image. It produces a grid of scores, where each score is the discriminator's confidence that a patch from the reconstructed image is similar to the corresponding patch on the ground truth image [44].

As demonstrated in Figure 2.9, in the MRI-to-CT synthesis task, the generator produces fake samples based on x (MR images), while the discriminator model observes x (MR images) and either the generated sample, $y' = G(x)$ (sCT) or the real sample y (CT).

Common Issues in GANs

There are several common issues related to GANs' performance and stability, which are still

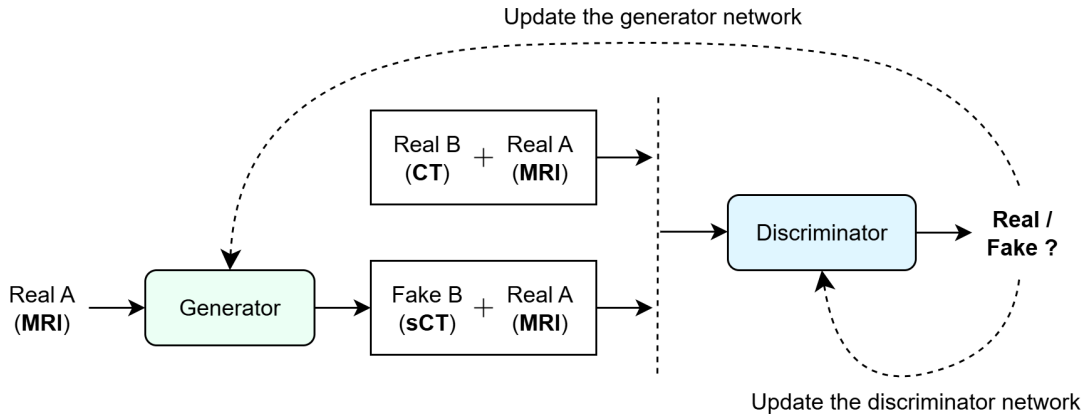


Figure 2.9: Simple cGAN training scheme in the case of MRI-to-CT synthesis task.

areas of active research. One significant obstacle is the **gradient vanishing problem**, which occurs when the D becomes too good at distinguishing fake images from real ones, causing the value function to saturate ($D(x,G(x,z))) \rightarrow 0$). This makes it difficult for the G to learn, as the gradients of the G loss approach zero ($log(1-D(x,G(x,z))) \rightarrow 0$). Even if the D becomes highly effective, there is no guarantee of reaching an optimal G.

Secondly, **mode collapse** happens when the G repeatedly produces a limited set of images with little to no diversity. This occurs because the G overfits the current D, producing outputs that the G knows will be classified as real. Assuming that the D is consistently fooled by this synthetic set and fails to adapt, the G may become stuck, generating a narrow range of outputs. Consequently, this feedback loop reduces variability in the generated outputs, as the G repeatedly produces highly similar images.

Finally, GANs frequently **fail to converge**. As the G improves, the D finds it increasingly difficult to differentiate between real and generated data, eventually reaching random guessing with 50% accuracy. This leads to weaker feedback from the D, which can cause the GAN to diverge if training continues, as the G relies on feedback that becomes less meaningful. As so, GAN convergence is typically unstable and brief [40].

Chapter 3

State of the Art

This chapter reviews the previous and current state-of-the-art approaches to generate sCT scans and sets the stage for the work to be undertaken in this thesis. Firstly, the rationale behind generating sCT from MRI data will be discussed. Following this, on the same section 3.1, the models designed for this purpose are presented. The section 3.2 provides a compilation of studies focused on producing CT-like images from MRI using Deep Learning (DL) techniques. Finally, the chapter concludes with a summary of the most robust DL architectures discussed in section 3.3, highlighting the ones that generate the most accurate sCT scans.

3.1 Generation of sCT scans

Recent studies have explored MR-only treatment planning for RT, in order to simplify this process by eliminating the need for separate CT scans, which would otherwise require a co-registration of the two modalities. In the absence of the CT exam, the ambiguities associated with co-registration are eliminated, and the workflow becomes faster [3]. Therefore, MR-only reduces patient radiation exposure, minimises discomfort experienced, and decreases planning-related costs and time for medical staff [4] [5].

Nevertheless, there are certain challenges of using MR-only RT for dose calculation, such as the lack of information on tissue attenuation properties provided only by CT [4]. To overcome these limitations, the literature has proposed the generation of sCT images from MRI for RT and dose planning [4]. These synthetic, or Pseudo-computed Tomography (pCT) images are created in HU and derived directly from MRI data [4]. Although this solution enables a successful MR-only RT planning, some constraints remain, such as limited access to MRI centres, regulatory obstacles, and the need for advanced sCT technology to ensure clinical applicability [3]. To accomplish the objective of generating sCT images, several approaches have been explored, which can be divided into four categories: Bulk-density Method (BDM), Atlas-based Method (ABM), Patch-based Method (PBM), and Deep Learning Method (DLM) [5].

3.1.1 Bulk-density method

The BDM involves segmenting MRI images into several classes, such as air, soft tissue, and bone, either manually by an expert or through an automatic process. Once segmented, each class is assigned a fixed density value from the training data that includes corresponding CT scans. These fixed density values are then used to approximate the electron densities needed for dose calculations and generate the corresponding sCT [5]. However, this method is heavily dependent on the operator due to the required manual segmentation, tedious, time-consuming, and does not consider tissue inhomogeneity [4].

3.1.2 Atlas-based method

The ABM was proposed in 2015 by Dowling et al. [6]. It consists of aligning new target MRI data with a MRI atlas that contains known MRI-CT correlations. This alignment is achieved using non-rigid registration, which uses one or more co-registered atlas pairs [5]. Non-rigid or deformable registration refers to the process of aligning or transforming data to bring them into a common spatial reference frame, taking into account anatomical or structural differences between datasets. This presupposes more complex and flexible adjustments compared to rigid registration [7]. A small neighbourhood (or patch) around each voxel in the target MRI is chosen and compared to the corresponding region in the registered MRI atlases. The differences between these patches is used as normalised weights, which are applied to the corresponding locations in the registered CT atlases to generate each voxel in the resulting sCT [5]. The drawbacks include reduced robustness when dealing with significant anatomical variations, dependence on an accurate co-registration process, and the need for substantial computational resources for pairwise registrations [4].

3.1.3 Patch-based method

While the PBM shares similarities with the ABM regarding the comparison between MRI data and reference information for sCT generation, there are certain differences in how this reference data is used. In 2018, Largent et al. [8] introduced the PBM, a machine learning-based approach involving four steps. It begins with rigid or affine registration of inter-patient MRI scans. Then, features are extracted from the registered MRI to obtain spatial, textural, and gradient information. After this step, the method selects the training patches most similar to the target MRI patches. Finally, a multipoint-wise aggregation is performed to generate the corresponding sCT patches [5]. Despite its effectiveness, the method has disadvantages, such as the calculation time and imprecise inter-patient registration [4] [5].

In summary, the ABM uses an atlas of paired MRI-CT scans and applies deformable registration to adjust and align the whole anatomy of the target MRI with the atlas. Meanwhile, the PBM relies on local feature extraction by selecting patches from the rigidly registered MRI, and identifying the training patches with the highest similarity to those of the target MRI to generate the sCT [4].

3.1.4 Deep learning method

DLMs have become increasingly popular in the generation of sCT due to their superior performance and promising deep learning-based architectures over traditional methods (i.e. ABM) [9]. The application of DLMs for the generation of sCT from pelvis MRI data has been explored in at least 18 studies [9]. The growing adoption is attributed to the advantages of DLMs. These models contain multiple layers that learn complex data representations, allowing them to effectively understand the relationships between HU (or CT values) and MRI intensities. Furthermore, some models do not require complex inter-patient registrations and generate sCT images very fast [4] [5].

3.2 Related work

Table 3.1 presents a collection of studies that explore DLMs approaches for MRI-to-CT generation in the pelvis or prostate region. For each study, a summary and an explanation of the workflow are provided in the following paragraphs. The architectures, loss functions, and metrics employed in the selected studies are further detailed in Table 3.2, Table 3.3, and Table 3.4, respectively.

The generation of sCT from MRI scans involves the use of different loss functions to monitor the model's performance during training. Least Absolute Deviations (L1) and Least Square Errors (L2) loss functions measure differences in image pixels or voxels, while Kullback-Leibler Divergence (KLD) loss function measures the differences between probability distributions. The loss functions based on the Perceptual Loss (PL) functions utilise pre-trained networks to extract and compare features in the images, resembling the human visual system. Finally, a more advanced loss function is the Content and Style Representation for Enhanced Perceptual Synthesis (CREPS), which uses a specific network to extract features focused on anatomical structures and texture.

A comprehensive set of metrics is key to evaluating the model's performance and sCT suitability for RT planning. This is achieved by assessing both image quality and dosimetric accuracy. Imaging metrics, including Mean Error (ME), MAE, PSNR, SSIM, MS-SSIM, Normalized Cross-correlation (NCC), and Pearson Correlation Coefficient (PCC), measure the pixel-wise, structural similarity and linear correlation between images. Dose metrics, including DVH differences and gamma pass-rates are essential for ensuring that the calculated radiation dose distribution on the sCT is clinically acceptable.

Largent et al. [45], compared and evaluated DLMs with different loss functions and HP to manage model training, including the U-Net (with loss functions L2 and single-scale PL) and the GAN (with loss function L2, single-scale PL, multiscale PL and weighted multiscale PL), alongside the PBM for sCT generation in MRI-based prostate cancer dose planning. The DLMs were trained on anatomically paired 2D axial slices of T2-weighted MRI and CT images. For all generation methods, 39 patients' MRI and CT images were utilized and randomly split into 3 training/validation cohorts as pictured in Table 3.1. The evaluation was performed based on

imaging endpoints, and dose endpoints. Concerning imaging endpoints, GAN L2 and U-Net L2 achieved the lowest MAE (<34.4 HU), and near zero ME for prostate CTV, with GAN L2 achieving lower uncertainty in bone image generation compared to U-Net. DLMs outperformed PBM in imaging accuracy, dose uncertainty, and computation time for sCT generation. The DVH points from sCTs showed no difference compared to those from real CT, when using GAN L2 and U-Net L2. Thus, the dose uncertainties computed with sCT scans generated by either method were small and clinically insignificant. However, the study had limitations, such as a small dataset size, exclusive use of T2-weighted MRI sequences, and geometric uncertainties originated from non-rigid registration. Despite these limitations, DLMs, particularly U-Net L2 and GAN L2, demonstrated promising clinical integration into MRI-only workflows.

The original paper by Bahrami et al. [9] aimed to investigate the performance of several widely used DL architectures for sCT generation from pelvis MRI, including, Efficient Convolutional Neural Network (eCNN), U-Net, V-shaped Convolutional Neural Network (V-Net), Residual Network (ResNet), and GAN. The results of these DL models were compared against an ABM, serving as a baseline for performance assessment. As shown in Table 3.1, even though the Cross-validation (CV) scheme differed between the DLM and ABM, both utilised co-registered 3D T2-weighted MRI-CT pairs. Regarding loss functions, the models performed better with L2-norm loss function, while the GAN's discriminator used MAE loss. eCNN, V-Net, and ResNet demonstrated lower noise levels, and the sCT produced resembled the ground truth CT images. Among these, eCNN obtained the lowest MAE (26.03 ± 8.85 HU) and ME (0.82 ± 7.06 HU), possibly attributed to its new building block structure, which enables a faster convergence rate during training, and an effective update of the parameters. Additionally, the highest PCC metric in the pelvic region was yielded by ResNet (0.91 ± 0.02) and eCNN (0.93 ± 0.05). In this study, the GAN method showed inferior performance due to the limited dataset size used. Nevertheless, all DLMs outperformed the ABM, which had the highest MAE (226.09 ± 85.07 HU). In the end, eCNN and ResNet demonstrated clinically acceptable performance with tolerable quantification errors.

A novel approach used the Pixel-to-pixel (Pix2Pix) architecture, a type of cGAN, to generate and optimise sCT images for MRI-only prostate RT planning [5]. The study aimed to test multiple generator architectures, loss functions (L1, L2, KLD and PL), and HP to improve imaging accuracy. Additionally, it compared Pix2Pix with five alternative methods: BDM, ABM, PBM, U-Net, and GAN. The Pix2Pix model performance achieved the lowest MAE for the whole pelvis (29.5 HU) using ResNet 9 blocks paired with PL between sCT and real CT images. This model surpassed similar cGAN methods used in prior studies, as confirmed by the Wilcoxon test (p-value$<$ 0.5, indicating that the paired samples have statistically significant differences), with a mean gamma passing rate of 99.4% (criteria: 1%/1mm, dose threshold: 10%). Compared to previously published sCT generation methods (BDM, ABM, PBM or DLM) with the same dataset, the Pix2Pix exhibited lower MAE of 29.5 HU. The cGAN architecture showed low dose uncertainties and fast calculation time, establishing itself as clinically acceptable. Nonetheless, its dose errors were similar to those of U-Net and GAN. The authors acknowledge challenges that compromised the study's evaluation metrics, particularly rectal gas variability, including

differences in gas pockets between MRI and CT images.

Texier et al. [46] proposed a 3D cGAN method to generate sCTs for RT dose calculation. The approach included both unsupervised (with paired and unpaired data) and supervised (using paired data) strategies. Supervised methods depended on CT/MRI registration precision and were typically limited to data from a single centre. To address this limitation, the study focused on enhancing the model's robustness through multicentre data, making it the only study approaching this concern among those presented in Table 3.1. In addition, the goal was to reduce dependence on registration. Although both supervised and unsupervised methods utilised cGAN built upon the Pix2Pix network backbone, the incorporation of an innovative CREPS loss was made through unsupervised training. This novel PL was formulated using the ConvNext-tiny network, which outperformed other image recognition methods. The results of this paper were promising, as multicentre data proved to accurately produce sCT images through unsupervised learning, thus eliminating the need for CT/MRI registration. This conclusion is supported by the finding that, the unsupervised paired network demonstrated a MAE of 27.5 ± 23.1 HU, similar to the results obtained with supervised architecture. Regarding dose calculation, performance was equivalent across architectures, with all models providing gamma pass rates above 94% (criteria: 1%/1mm, dose threshold: 10%), ensuring clinically acceptable results.

To improve RT planning by eliminating the need for CT scans and reducing errors associated with image registration, Pan et al. [47] proposed the first MRI-to-CT Diffusion Denoising Probabilistic Models (MC-DDPM) approach, a 3D transformer-based model developed to generate sCT scans matching MRI anatomy. The method involves two processes: a forward process, where Gaussian noise is incrementally added to real CT scans to create noisy scans, and a reverse process, where the noisy CT scans are denoised with a Shifted-window Transformer V-net (Swin-Vnet). This denoising process is conditioned on MRI data from the same patient to reconstruct noise-free CT scans. The approach was evaluated on a prostate dataset using T1-weighted MRI-CT paired data. For the prostate, the high quality sCT images was achieved with MC-DDPM, with a MAE of 59.953 ± 12.462 HU, PSNR of 26.920 ± 2.429 dB, SSIM of 0.849 ± 0.041, and NCC of 0.948 ± 0.018. These results showcase statistically significant improvements in evaluation metrics over competing networks according to the Student's paired t-test. Therefore, the MRI-to-CT Diffusion Denoising Probabilistic Models (MC-DDPM) produced images of higher quality compared to GAN and traditional Diffusion Denoising Probabilistic Models (DDPM). Even so, this exceptional performance comes at the expense of increased computational cost.

Table 3.1: Summary of methods and main results from studies on MRI-to-CT synthesis. The studies were conducted on prostate/pelvic regions using deep-learning-based approaches.

Study	Approach Used	Train/Test Dataset Size	Technical Evaluation	Brief Summary of Findings
[45]	**U-Net GAN PBM**	39 patients: 3 train/validation cohorts (25/14, 25/14, 25/11)	MAE, ME, Mean DVHs, Gamma pass rate, Mean gamma	DLMs demonstrated potential for MRI-based prostate dose planning. GANs, in particular, produced more realistic images.
[9]	**eCNN U-Net V-Net ResNet GAN Atlas-based**	20 patients: DLM 5-fold CV (15/5), Atlas-based leave-one-out CV (19/1)	MAE, ME, PCC, SSIM, PSNR	eCNN and ResNet demonstrated strong performance among DLMs with clinically acceptable quantification errors.
[5]	**cGAN (Pix2Pix)**	39 patients: 3 train/test cohorts (25/14, 25/14, 25/11)	MAE, ME, DVH, 3D Gamma pass rates	The Pix2Pix architecture with ResNet-9 blocks and PL shows promise for MRI-based prostate cancer dose planning.
[46]	**3D unsupervised and supervised learning cGAN**	C1: 39 Patients C2: 30 patients C3: 30 patients 70% (train); 10% (validation); 20% (test)	MAE, ME, PSNR, Gamma pass rates, DVH	Unsupervised learning using multicentre data produced accurate sCTs without the need for CT-MRI registration.
[47]	**MC-DDPM**	28 patients: 20 MRI-CT (train); 2 MRI-CT (validation); 6 MRI-CT (test)	MAE, PSNR, MS-SSIM, NCC	MC-DDPM showed significant improvements for prostate sCTs compared to other networks.

Table 3.2: Definition of several architectures used to generate CT-like images.

Network	Definition
eCNN [9]	Based on the **encoder-decoder architecture of U-Net**, the eCNN model was designed to extract discriminative image features from MRI. It replaces convolutional layers with a **building structure** that **transfers information from upper to lower layers**, preserving important details.
U-Net [9][45]	The U-Net is an **encoder-decoder network** that **downsamples** the input MRI to extract complex features and **upsamples** them to reconstruct the pCT. **Skip connections** transfer high-resolution features from the encoder to the corresponding level of the decoder.
V-Net [9][48]	**3D network similar to U-Net**, with symmetric compression (encoder) and decompression (decoder) paths. These paths consist of multiple stages with **convolutional layers and residual connections**, enabling it to capture complex features.
ResNet [9][49]	ResNet builds upon **residual blocks**, which directly **adds the input of a block to its output** by a **shortcut connection** that skips one or several layers, addressing the **vanishing gradient problem**.
GAN [50]	GAN is composed of a **generator (G)** and a **discriminator (D)**. The G generates realistic sCT samples from noise, which are then combined with real data and passed into the D. D estimates the probability of each sample being real or fake. Through **adversarial training**, D's feedback guides G in optimising its parameters for the next generation.
cGAN [5][50]	The cGAN is a **GAN variant** that incorporates additional information, such as the input image modality, as a **conditional variable** into both the generator and discriminator. The output sCT is conditioned on the input MRI, enabling the model to learn **image-to-image translations**.
Pix2Pix [4][5]	Pix2Pix is a **cGAN variant** for **high-resolution image-to-image translations** using typically **U-Net as the generator** and **PatchGAN as the discriminator**. The PatchGAN discriminator evaluates the difference between fake and real images by analysing local patches.
DDPM [51]	The DDPM involves a **forward diffusion process**, where a predefined noise distribution is incrementally added to the image, and a **backward denoising process**, where a neural network predicts the amount of noise added at each step and progressively subtracts it from the noisy image.
MC-DDPM [3]	The MC-DDPM stands for MRI-to-CT DDPM, a **transformer-based approach** that generates pCT from MRI, by **conditioning the denoising process** on MRI inputs.

Table 3.3: Definition of the most common loss functions utilised to generate sCT images.

Loss Function	Definition	Description
L1 [45]	$\mathrm{L1} = \sum_{i=1}^{n} \lvert y_{\mathrm{gt}} - y_{\mathrm{pred}} \rvert$	Minimizes the error through the sum of the absolute differences between a prediction and the ground truth.
L2 [45]	$\mathrm{L}_{\mathrm{Network}}(I,C) = \lvert\lvert C - Network(I) \rvert\rvert_2^2$	Minimizes the differences between CT and sCT voxels.
KLD [52]	$\mathrm{KLD(p\lvert\lvert q)} = \sum_{i=1}^{n} p(x_i) \log(\frac{p(x_i)}{q(x_i)})$	Calculates the difference between the predicted and true class probability distributions.
PL [45]	$\mathrm{L}_{\mathrm{Network}}(I,C) = \lvert\lvert VGG(C) - VGG(Network(I)) \rvert\rvert_2^2$	By mimicking the human visual system, PL compares CT and sCT images similar features using a pre-trained VGG network (or other), rather than pixel-wise differences.
Multi-scale PL [45]	$\mathrm{L_G}(I,C) = \frac{1}{card(S)} \sum_{i \in S} \lvert\lvert VGG_i(C) - VGG_i(G(I)) \rvert\rvert_2^2$	By using multiple layers of a pre-trained VGG network (or other), it compares features of the images at different scales.
Weighted Multi-scale PL [45]	$\mathrm{L_G}(I,C) = \frac{1}{card(S)} \sum_{i \in S} w_i \lvert\lvert VGG_i(C) - VGG_i(G(I)) \rvert\rvert_2^2$	It uses multiple layers of the VGG network (or other), with each layer assigned a different weight. Layers that yield the lowest MAE are given more importance.
CREPS [46]	CREPS relies on ConvNext-Tiny network for tailored feature extraction to capture anatomical structures (content) and texture information (style).	An innovative PL function that focuses on separating the content and style of the image to synthesize CT images from MRI.

I - MRI; C - corresponding CT; Network - Neural Network used to generate I; $||\ ||_2^2$ - L2 norm or Euclidean norm (square root of the sum of the squares of its components); $p(x_i)$ is the true probability of class x_i and $q(x_i)$ is the predicted probability of class x_i.; VGG (Visual Geometry Group) - output of a VGG16 convolutional layer. The VGG16 network (16 convolutional layers) was selected due to being commonly used for PL computation for various tasks such as image deblurring and super-resolution.; $S = \{2,5,7,10,13\}$; $G(i)$ - pCT produced by the generator; VGG_i - i^{th} VGG16 convolutional layer;$w_i = e^{-(\mathrm{MAE}_i(C,G(I))}$ with MAE_i being the mean absolute error between the CTs and pCTs generated by the GAN.

Table 3.4: Definition of imaging and dosimetric endpoints. DVH and its difference, as well as gamma pass rate are dose metrics. The remaining are imaging metrics.

Metric	Definition	Ideal value
ME	$\mathrm{ME} = \frac{1}{N}\sum_{i=1}^{N} pCT_i - CTi$	0 HU
MAE	$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \lvert pCT_i - CTi\rvert$	0 HU
PSNR	$\mathrm{PSNR} = 10\log_{10}\frac{Q^2}{MSE}$	Maximum dB
SSIM [53]	$\mathrm{SSIM(x,y)} = [l(x,y)]^{\alpha}\cdot[c(x,y)]^{\beta}\cdot[s(x,y)]^{\gamma}$ (with $\alpha=\beta=\gamma=1$) $\mathrm{SSIM(x,y)} = \frac{(2\mu_x\mu_y+C_1)(2\delta_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\delta_x^2+\delta_y^2+C_2)}$	1
MS-SSIM [53]	$\text{MS-SSIM}(x,y) = [l_M(x,y)]^{\alpha_M}\cdot\prod_{j=1}^{M}[c_j(x,y)]^{\beta^j}\cdot[s_j(x,y)]^{\gamma^j}$	1
NCC	$\mathrm{NCC} = \frac{1}{N}\sum_{x,y,z}\frac{(I_{CT}(x,y,z)-\mu_{CT})(I_{pCT}(x,y,z)-\mu_{pCT})}{\sigma_{CT}\sigma_{pCT}}$	
PCC [9]	$\mathrm{PCC(CT,pCT)} = \frac{\sum_{i=1}^{N}(CT(i)-\overline{CT})(pCT(i)-\overline{pCT})}{\sqrt{\sum_{i=1}^{N}(CT(i)-\overline{CT})^2}\sqrt{\sum_{i=1}^{N}(pCT(i)-\overline{pCT})^2}}$	1
DVH	Dose distribution delivered to different volumes of a tissue/organ, used to assess how closely sCT matches real CT dosimetry.	-
DVH difference	Dose difference on specific DVH points for a given structure.	0Gy/0%
Gamma pass-rate	This metric evaluates the agreement between sCT and reference CT dose distribution using gamma analysis.	100%

Table adapted from [4]. Abbreviations: N - number of voxels; Q - intensity of i^{th} voxel in the sCT and reference CT images; x - reference CT; y - sCT; μ_x and μ_y - mean values of x and y; δ_x^2 and δ_y^2 - variances of x and y; δ_{xy} - covariance of xy $C_1 = (k_1Q)^2$; $C_2 = (k_2Q)^2$ with k_1 and k_2 being constants; $l(x,y)$, $c(x,y)$ and $s(x,y)$ represents luminance, contrast and structure image components, respectively; M - Scale; j - index of the scale; I_{CT} and I_{sCT} - HU values of the reference CT and sCT; μ_{CT} and μ_{sCT} - mean intensity values of the reference CT and sCT; σ_{CT} and σ_{sCT} - standard deviations of the reference CT and sCT; $\overline{\text{CT}}$ and $\overline{\text{sCT}}$ - means of reference CT and sCT.

3.3 Summary

Analysing the studies presented in Table 3.1, it is evident that U-Net and cGAN architectures are the most commonly used approaches for MRI-to-CT generation. However, focusing on the main findings of each paper, it is notable that both GAN-based methods and diffusion models demonstrate the ability to accurately generate sCT scans.

According to the literature [3], cGANs and the integration of ResNet blocks in DL architectures are worth exploring as a research avenue. Unlike the unconditional GAN architecture, the cGAN condition both the generator and the discriminator on an input MRI, enabling the generation of a corresponding output CT-like image. In theory, during the training of cGAN the loss is learned, penalising any structural differences between the output (sCT) and the target (CT) [39]. To put in another words, it would be meaningless to translate MRI into CT if the resulting CT-like image does not accurately preserve the anatomical details of the individual who performed the input MRI. To handle this, attention mechanisms are incorporated into architectures, in order to focus on important MRI anatomical features that play a significant role in the electron density of the CT image [3].

As reported by Table 3.1, a group of commonly used metrics evaluates the similarities between sCT and CT images. These include MAE, ME, PSNR, and SSIM, which focus on estimating the precision of the pixel level. For evaluating the calculated dose on sCTs and comparing it with dose calculations on CTs, most studies employed DVH, DVH differences, and gamma pass-rates. These findings align with the literature presented by Bahloul et al. [3].

Despite significant advancements in generating realistic sCT from MRI, limitations such as small dataset sizes, data quality sensitivity, and high computational costs still hinder the robustness and generalisability of these DLMs. Even though, novel DL architectures for sCT generation to MR-only RT remain a work of research due to its promising future to improve patient-focused treatment delivery and planning [3].

As such, my research is going to undertake the specific issues associated with MRI-to-CT generation. By tackling **small dataset sizes** and quality sensitivity, through the use of a **larger**, **high-quality**, and **diverse** dataset, along with **developing an innovative strategy** to create a robust and high-level DLM, this work will contribute to advancing cancer RT planning based solely on MRI, offering a more clinically reliable solution.

Chapter 4

Material and Methods

This chapter details the methodology employed, beginning with a description of the two datasets used and their acquisition (section 4.1), followed by the processing steps, such as data augmentation (section 4.2). Subsequently, an overview of the architectures utilised in section 4.3 is provided, along with the metrics used to evaluate generated images in section 4.4.

4.1 Dataset description

In this work, two publicly available datasets from **Synthesising computed tomography for radiotherapy Grand Challenge (SynthRAD)** were used. This challenge aims to develop sCT generation methods from both MRI (challenge Task 1) and Cone Beam Computed Tomography (CBCT) (challenge Task 2) by providing the first large-scale public multi-centre dataset along with the evaluation metrics. The **SynthRAD** deep learning challenge followed a similar structure in both selected editions, *SynthRAD2023* (pelvic data from centres A and C) and *SynthRAD2025* (abdominal, thoracic, and head and neck data from centres A, B, C, and D), which are explained in more detail below. Only *Task 1* was considered, as it addresses MRI-to-CT conversion for MR-only RT.

Concerning patient characteristics, such as tumour type and staging, these were not available in either dataset to preserve patient privacy. However, the age and sex for each patient were available in SynthRAD2025, for abdominal, thoracic, and head and neck regions obtained in centre A, and for the head and neck images from centre D (except for 2 patients). In the centre A set, female (F) images outnumbered male (M) images in the thoracic (F-65, M-26) and head and neck (F-64, M-27) regions, but the opposite was observed in the abdominal region (F-26, M-37). Centre D dataset consisted predominantly of male cases, with 47 images from male cases compared to 16 images from female cases. For the thoracic region, patients' ages ranged from 36 to 88 years, with a mean age of approximately 63 years. In the abdominal region, the youngest patient was 6 years old and the oldest was 84 years, with an average age of 65 years. The head and neck region in centre D had a mean age of 61 years, ranging from 43 to 81 years, and in centre A the mean age is approximately 65, with a minimum age of 29 years and a maximum of 87 years. SynthRAD2023 reported a general overview of patient demographic data, in which

72.6% were male cases and 27.4% female, making this a predominantly male dataset due to the inclusion of prostate patients, and the mean patient age was 65, with ages ranging from 3 to 93 years.

In terms of the preprocessing pipeline, the publicly available datasets SynthRAD2023 and SynthRAD2025 were both pre-processed by the challenge authors to facilitate synthetic CT generation. The automatic pre-processing pipeline included resampling the images from both datasets to the same voxel spacing, data conversion from DICOM files to compressed NIfTI (.nii.gz) for SynthRAD2023 and to compressed MetaImage format (.mha) for SynthRAD2025. Additionally, to address misalignment between the dual-modality images, both datasets underwent rigid registration of MRI and CT scans using the Elastix framework [54]. Moreover, each case contained a body outline segmentation as a binary mask, crucial for the training process and metric calculations during validation. Finally, both datasets cropped the images to remove the background and minimise the file size. Furthermore, the facial features of the HN images (SynthRAD2025 dataset) were blended into the background through a process called defacing, to ensure patient anonymity.

Although both datasets are multi-centre, which contribute to their heterogeneity, this also introduces limitations due to differences in MRI sequences and CT reconstruction methods across centres. Moreover, within each centre, scanners (model and manufacturer) and imaging protocols differ, compromising the consistency of image characteristics for the same dataset. More details on acquisition parameters are documented in [55] and [56].

4.1.1 SynthRAD2023

The *SynthRAD2023* dataset ([55]) contains co-registered pairs of planning MRI-CT scans and body masks for the male and female pelvic regions acquired at two different centres, as pictured in Figure 4.1. Moreover, to prevent the generation models from relying too much on the mask itself, the final mask is dilated, including the surrounding air.

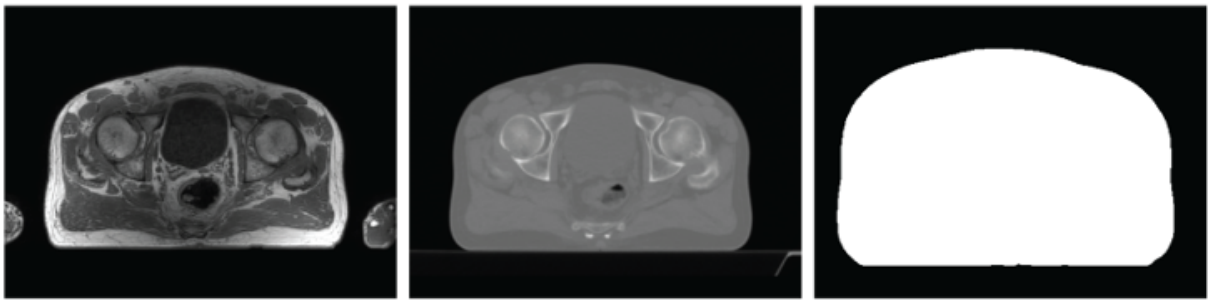


Figure 4.1: Example of pelvis image from SynthRAD2023. MRI (left), CT (middle), and the associated dilated body outline (right). Image from [55].

Even though the dataset includes 180 subjects for training, 30 for validation and 60 for testing, only the training folder was used. The reason for this was the absence of CT ground truth images in the validation and test folders, which were vital for the experimental phase of the work. The original training folder was split into train (90%) and test subsets (10%). Of the

180 patients who underwent MRI and CT exams, 120 acquired pelvic images in Centre A and 60 in Centre C (from two Dutch university medical centres, anonymised).

Data acquisition

Regarding the acquisition of pelvic MRIs, centre A used a Philips scanner with two field strengths (1.5 T and 3 T) and employed a spoiled T1-weighted gradient echo sequence, while centre C used a Siemens scanner with a single 3 T field strength, and acquired a T2-weighted fast spin echo sequence. CT scans were also acquired using different scanners: centre A used two scanner types, one from Philips and the other from Siemens, whereas centre C used solely a Philips scanner. The slice thickness ranges from 1.5-3mm for images in centre A and 2-3mm in centre C. The MRI-CT pairs differ in dimensions, but were resampled by the challenge authors to a uniform voxel spacing of 1x1x2.5mm.

4.1.2 SynthRAD2025

The *SynthRAD2025* dataset ([56]) provides co-registered MRI-CT pairs from three different anatomical regions, including patients with HN, TH, and AB cancers, along with their body masks. Various examples are illustrated in Figure 4.2, Figure 4.3, and Figure 4.4. These were collected from four European university medical centres (radiation oncology departments), out of the following five (the fifth centre was used for challenge Task 2, which was the CBCT-to-CT conversion): UMC Groningen, UMC Utrecht, Radboud UMC (Netherlands), LMU University Hospital Munich, and University Hospital of Cologne (Germany). The exact correspondence between these centres and the dataset's centre labels (e.g. centre B) is not provided.

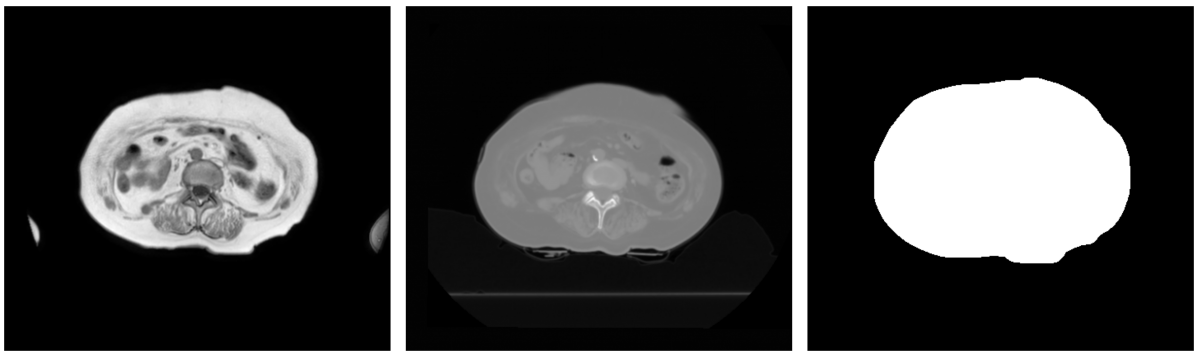


Figure 4.2: Abdominal region example from SynthRAD2025 dataset: MRI (left), CT (middle), and the associated dilated body outline (right).

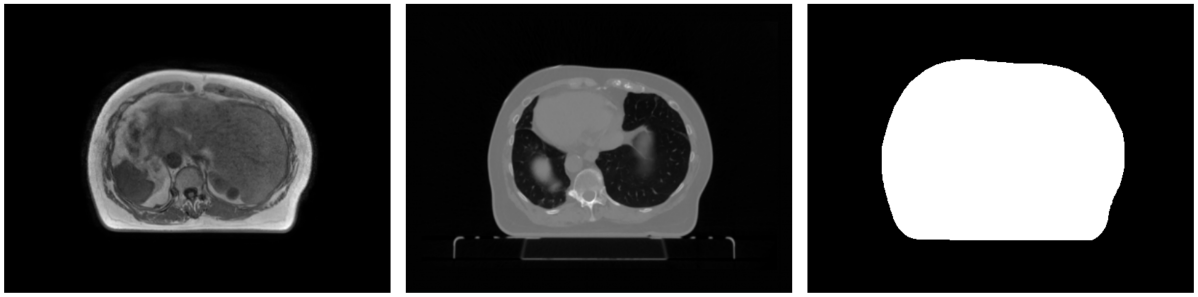


Figure 4.3: Thoracic region example from SynthRAD2025 dataset: MRI (left), CT (middle), and the associated dilated body outline (right).

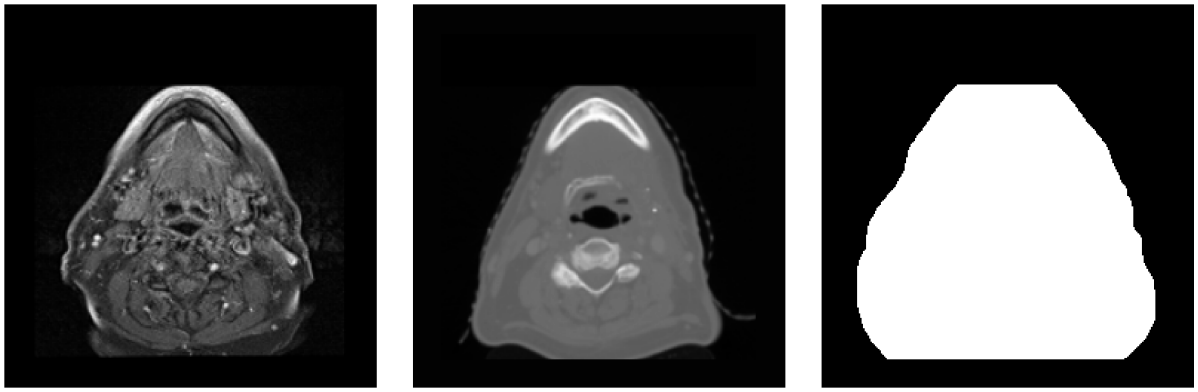


Figure 4.4: Head and neck region example from SynthRAD2025 dataset: MRI (left), CT (middle), and the associated dilated body outline (right).

Based on imaging protocols used in the different regions and centres, 890 pairs were acquired. In total, 340 MRI-CT pairs were from the head-and-neck region, 270 from the abdominal, and 280 images from the thoracic region. Considering that the SynthRAD Grand Challenge 2025 edition only released the validation dataset (10%) on June 1st of this year and the test dataset (25%) would be released in 2030, the training dataset (65%: 578 images), was the only one available for use, including 221 from HN, 175 from AB, and 182 from TH. Thus, the original training dataset was used in this work for both training and evaluating the model. Each case was assigned a unique patient ID (eg., 1ABA001), where the first digit was a task identifier (1), followed by a region identifier (HN, TH, or AB), a centre identifier (A, B, or C), and a three-digit patient pair number. A description of the number of cases at each centre and region is detailed in Table 4.1.

Table 4.1: The number of cases each centre (letter from A to D) provided for the three anatomical regions: HN, TH and AB. The division of images per centre/region is done by reserving 90% for training/validation and 10% for testing. Centre D serves exclusively as an external test and is never used for model training.

	Centre A	Centre B	Centre C	Centre D	Total
AB	65	91	19	0	175
TH	91	91	0	0	182
HN	91	0	65	65	221
Total	247	182	84	65	578

Data acquisition

Regarding acquisition protocols (Table 4.2 and Table 4.3), these vary according to the centre and in some cases within each centre. In certain centres, the MRI acquisitions were not performed in 3D. Exceptions included centre C (head and neck and abdominal) and centre B (abdominal and thoracic), where the acquisitions were in 2D. In particular, the centre C of the abdominal region included 2D and 3D acquisitions. Similarly to the SynthRAD2023 dataset, the dimensions of MRI-CT pairs vary across different pairs, but remain constant within an individual pair. Furthermore, all pairs have been resampled to a uniform voxel spacing of 1x1x3mm.

Table 4.2: Imaging parameters for CTs from all regions.

Region	Parameter	Centre A	Centre B	Centre C	Centre D
AB	Manufacturer	Philips	Toshiba	Philips and Siemens	-
	Slice thickness (mm)	2-3	3	2-3	-
TH	Manufacturer	Philips and Siemens	Toshiba	-	-
	Slice thickness (mm)	2-3	3	-	-
HN	Manufacturer	Philips and Siemens	-	Philips and Siemens	Siemens Healthineers
	Slice thickness (mm)	2	-	3	2

Table 4.3: Imaging parameters for MRIs from all regions.

Region	Parameter	Centre A	Centre B	Centre C	Centre D
AB	Manufacturer	Philips	ViewRay	Philips/Siemens	-
	Field strength (T)	1.5	0.35	1.5/3	-
	Sequence type	T1-weighted spoiled gradient-echo Dixon and T1-weighted radial fat-suppressed gradient echo	Balanced steady-state free-precession (T2/T1-weighted)	Spin-echo (T2-weighted) (12 patients)/Spin-echo and gradient recalled echo (T2/T1-weighted) (7 patients)	-
TH	Manufacturer	Philips	ViewRay	-	-
	Field strength (T)	1.5	0.35	-	-
	Sequence type	T1-weighted spoiled gradient-echo Dixon and T1-weighted radial fat-suppressed gradient echo	Balanced steady-state free-precession (T2/T1-weighted)	-	-
HN	Manufacturer	Philips	-	Siemens Healthineers	Siemens Healthineers
	Field strength (T)	3	-	1.5/3	3
	Sequence type	T1-weighted spoiled gradient-echo Dixon	-	T1-weighted turbo spin-echo	T1-weighted radio-frequency-spoiled gradient echo Dixon

4.2 Additional data processing

The entire set of transformations previously mentioned was executed by the challenge's authors, and from this point onward, the focus is on the additional transformations specific to this master's thesis.

Initial Processing

In both datasets, additional processing was necessary to prepare the images for training and evaluation with MONAI, which is a PyTorch-based framework originally developed by NVIDIA and King's College London, designed for deep learning in healthcare imaging. These processes ensured compatibility with the deep learning framework used throughout this work, PyTorch. For MRI scans, intensities were normalised between -1 and 1, given the fact that their voxel values are relative and not an absolute value. On the other hand, CT values represent quantitative HU measurements. Therefore, intensities were scaled within a specific range to correspond to the normalised range of [-1, 1]. The processing pipeline began with the following transformations:

1. 3D MRI, CT and mask volumes were loaded;
2. The first dimension was assigned as the channel dimension, to comply with MONAI's requirements. Thus, all volumes were converted into 4D tensors with shape [C, H, W, D], where C denotes the channel dimension (C=1 for grayscale);
3. The subsequent steps depended on whether the images were used for training or evaluation.
   - **Training process**:
     3.1. Mask slices with peak to peak intensities equal to zero (i.e. empty slices) were detected and removed, along with their corresponding slices in MRI-CT pairs.
     3.2. To guarantee meaningful content in the training samples, which were a result of random 256×256 2D crops in each slice, only those slices in which at least 10% of the pixels in the binary reference mask were non-zero were retained. All other slices were removed.
     3.3. The whole resulting MRI volume was normalised to intensities between -1 and 1.
   - **Evaluation process**:
     3.1. To preserve the image dimension, a copy of the original MRI image was created. In it, empty slices were eliminated and the minimum and maximum intensities from the filtered image were identified.
     3.2. After that, these minimum and maximum values were used to normalise the intensity of the original MRI volume between -1 and 1.
4. CT informative values between -1024 HU and 3000 HU were normalised to a range of $[-1, 1]$.

Further transformations were carried out only on the SynthRad2025 dataset, including decompressing the *.mha* files to enable their import into the matRad software (details below in subsection 4.4.3).

Technical implementation: The images were structured in a dictionary format to use the built-in functionalities of MONAI, particularly those based on dictionaries. *Compose* (MONAI transforms function), which allowed chaining a sequential series of callables, was used to apply the aforementioned transformations.

Pre-data-augmentation processing

Before applying data augmentation techniques to the **training** images, certain transformations were required to convert the 3D volumes into 2D slices.

1. **Spatial padding**: Zero-padded the images to a specific spatial size of 256×256. This was applied only if necessary, for instance, when the input dimension was smaller than 256.
2. **Random slice selection**: Keeping the original image height and width dimensions, the transformation was applied to randomly extract a single slice.
3. **Dimension squeeze**: In order to transform the 3D image, which had only one slice ([C,H,W,1]), to 2D, the depth dimension was squeezed ([C,H,W]).

During **validation**, only the 256×256 **Spatial padding** transformation was applied to its images.

Data augmentation

Several data augmentation transformations were applied as a regularisation technique to prevent overfitting the model on **training** images and to increase the diversity of the SynthRAD2023 and SynthRAD2025 datasets. These augmentations not only expanded the variability of the dataset but also improved the model's ability to handle realistic variations in clinical images. For instance, small rotations simulate variations in patient positioning during acquisition, while intensity transformations on MR images mimic possible artifacts. In the literature, spatial transformations, such as rotation (including rotation by 90^{o} angles), shearing, scaling, flipping and elastic deformation, are common data augmentation techniques for MRI and CT images [57]. The chosen values of intensity transformations and random affine transformation were determined through trial and error, avoiding significant variations in MRI intensities, and preserving the realism of all images. When appropriate, values provided by MONAI transformation examples were adopted [58] [59] [60]. All transformations had a 50% probability of being applied to the input (intensity transformation) or both to the input and output (spatial transformation). The sole exception was that of random elastic deformations, which, due to their time-consuming nature, had an application probability of only 30%.

Firstly, for each **spatial** transformation common to all MRI, CT and mask images, an explanation follows, listed in order.

1. **Random affine transformation**: Combined several spatial transformations that were

performed as it follows:

1.1. Randomly rotated within a range of -5º and 5º, along axis x;

1.2. Randomly selected the scale factor in the range of 5% smaller and 5% larger along axes x and y;

1.3. Randomly selected the shear factor from the range [-0.3, 0.3] radians, tilting the image along the x and y axes;

1.4. Any newly added voxels were filled with zeros;

1.5. Applied different interpolation modes according to each image type. The new values of CT and MRI were computed using bilinear mode, which uses the weighted average of the four nearest pixel values to preserve anatomical structure. Conversely, the mask interpolated values were calculated using nearest mode, which assigned the new pixel the exact value of its nearest neighbour, conserving the mask binary values [61].

2. **Random flipping**: The image was flipped along the y-axis.

3. **Random 2D elastic deformation**: Implemented a random elastic deformation with a distance between the control points of 20 pixels (spacing) for both x and y axes, a magnitude range of 1-2 pixels of displacement, and with the same interpolation modes as RandAffined.

4. **Random rotation by 90º**: New images were produced by randomly rotating them by 90º in the plane defined by the height and width axes (spatial axes 0 and 1).

Secondly, for each **intensity** transformation that imitates artifacts on real MRI, an explanation is listed in order below and illustrated in Figure 4.5.

1. **Random Gibbs noise**: One of the most common artefacts in MRI scans, the Gibbs artefact, was mimicked using a Gibbs noise filter with intensity parameter alpha, uniformly sampled between 0.6 and 0.8. The filter truncated the Fourier series, which originated the Gibbs artefact upon reconstruction of the MR signals into images.

2. **Random spike noise**: In order to simulate spike artefacts in MRI, the log-intensity of the interval (10,12) was sampled. These spikes are caused by large transient, localised, erroneous changes in k-space signal intensity during MRI data collection.

3. **Random bias field**: The bias field corruption was ensured by using a range of random coefficients of (0.02,0.05). This introduced intensity inhomogeneity of pixels in the MRI, simulating the realistic magnetic field perturbations.

4. **Random Gaussian noise**: This transformation added Gaussian noise with a mean of 0 and standard deviation of 0.05 to MRI images.

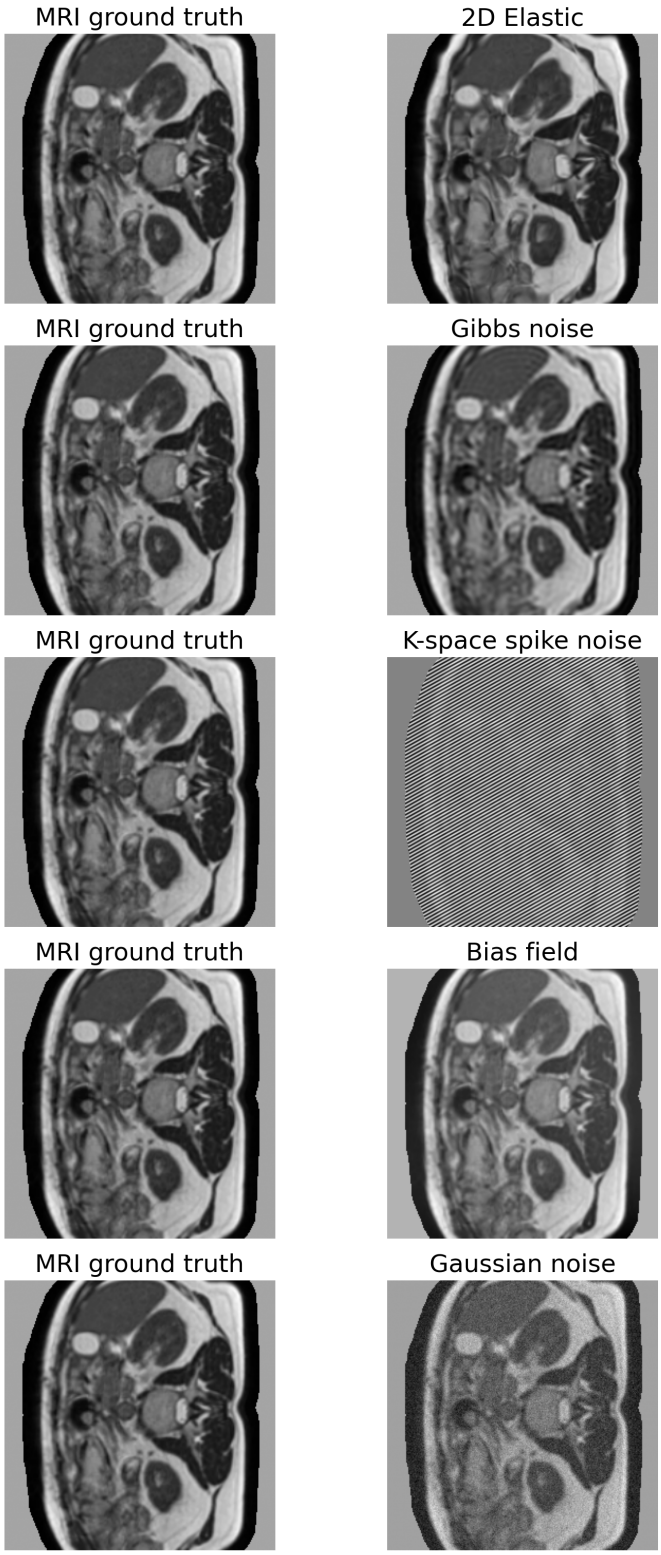


Figure 4.5: Example of the application of elastic transformation and intensity transformations to a single MRI slice from AB patient. The transformation parameters shown in this image differ from those used in this work and are intentionally exaggerated for visualisation purposes.

Post-data-augmentation processing

Finally, three spatial transformations were executed in MRI, CT, and mask. Their explanation is sequentially listed below.

1. **Random spatial crop**: After the application of all transformations, each image was cropped into a fixed region of interest (roi) of 256×256 size. The crop was taken from a random location. During training, one slice per training image was randomly selected and cropped. For validation, all slices were included and cropped to the specified size.
2. **Mask application**: The body outline binary masks were multiplied by the MRI and CT images, where the voxel values within the mask remained the same and those outside were converted to a constant value or zero.
3. **To tensor transformation**: Converted the images into tensor type.

4.3 MRI-to-CT generation

4.3.1 Autoencoder architecture

For a more simplistic architecture designed for image generation using DL, a U-Net-based generator was implemented. After HP optimisation, the generator (Figure 4.6) was modelled as a 256×256 U-Net with 8 layers. The architecture included a downsampling (encoder) path, an upsampling (decoder) path, and skip connections between their corresponding levels, preserving spatial details during the image generation process [39]. The input and output of the neural network had a single channel, as we use a single input and output modality (MRI and CT, respectively). In total, this network has 28,244,545 trainable parameters.

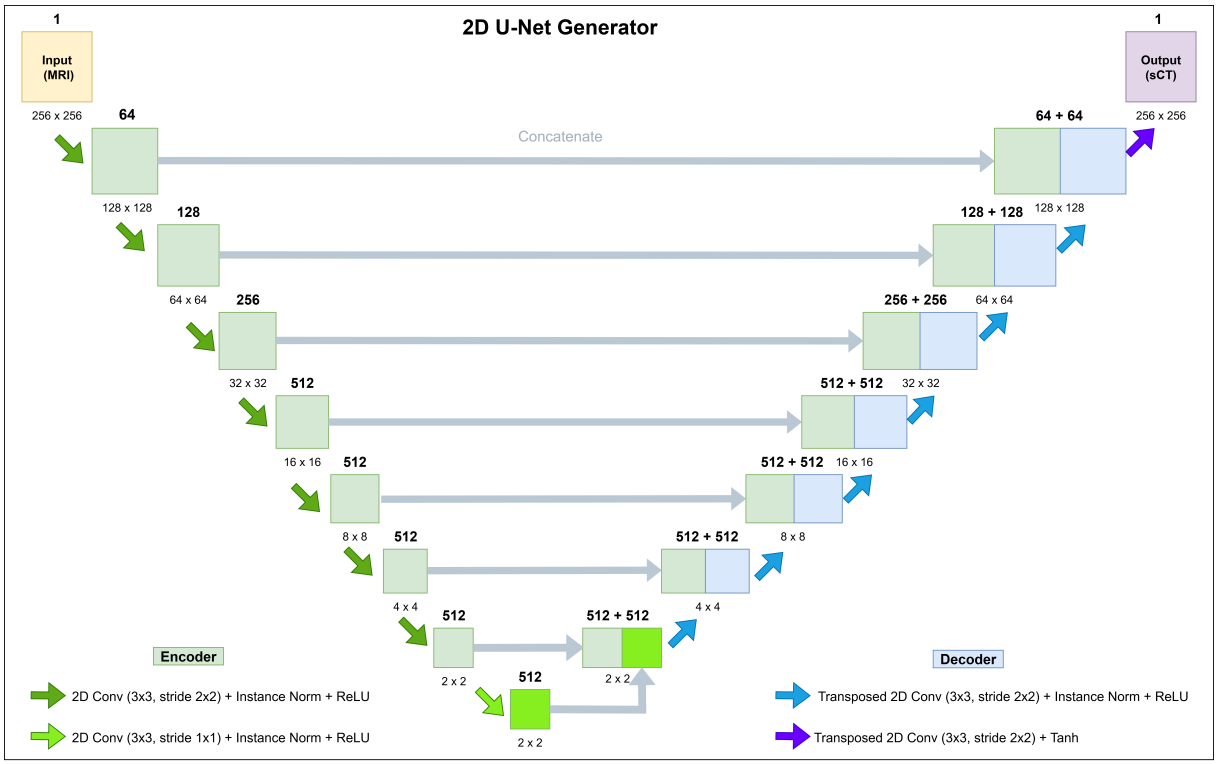


Figure 4.6: AE final architecture.

4.3.2 cGAN architecture

Generator

cGAN's generator architecture resembles that of the previously detailed AE. However, unlike the AE whose objective was the accurate reconstruction of the input images into another imaging modality, the cGAN generator was trained with an additional adversarial objective, which was to generate synthetic CT images that the discriminator was not capable of distinguishing from real ones. In terms of HP, the cGAN generator has 7 layers in its 256×256 U-Net (Figure 4.7), as opposed to the AE that has 8 layers.

Discriminator

In this work, a patch-based discriminator was used, inspired by the original image-to-image translation paper [39], to enable a classification based on different patches of size 32×32 with different local structures and textures, useful in medical imaging. The size of the patches was selected to enable the model to focus on local structures and detect small irregularities in those, similarly to [44]. First, the discriminator takes the pairs of patches of the MRI and corresponding real CT or sCT as inputs, concatenated along the channel dimension. Then, it downsamples the features, reducing the spatial dimensions while increasing the number of channels, through a series of 4 convolutional blocks.

Since the output of the convolutional blocks is not a single scalar prediction per patch and image, additional steps were required at the final stage of the discriminator architecture. The

output of the last convolutional block is first flattened across spatial dimensions, and a max pooling operation is applied to select the maximum value per feature channel. This results in a single vector representation of each image, which is subsequently passed through a set of fully connected layers with activation functions (the classification layers) to produce a final scalar prediction.

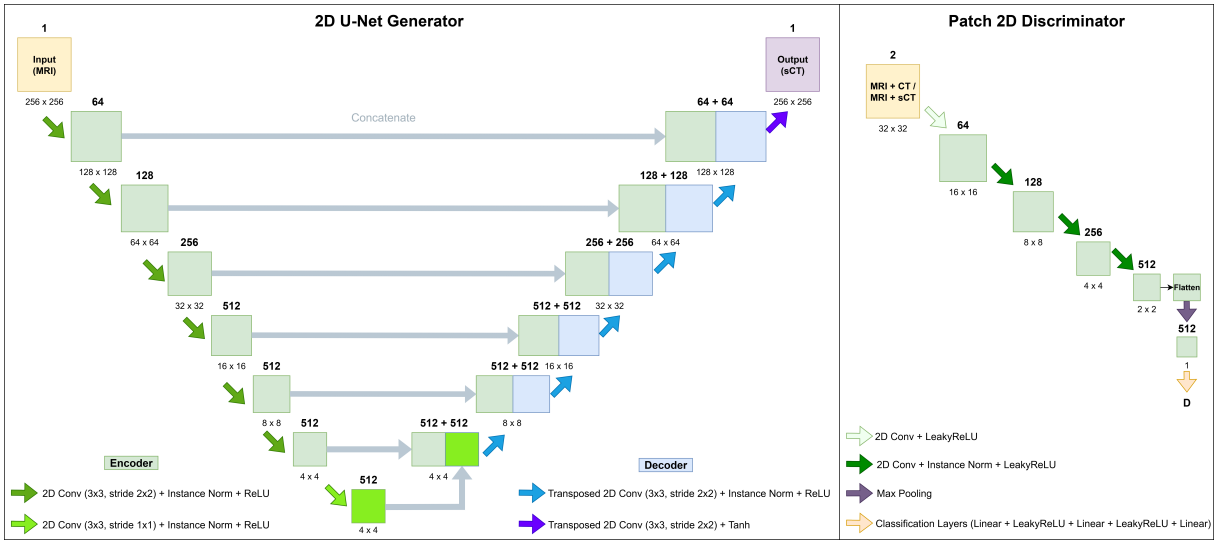


Figure 4.7: cGAN final architecture.

In contrast to what was implemented in [39], a sigmoid activation was not used as a final layer. Binary classification tasks commonly use the sigmoid activation to output probabilities in the range between 0 and 1, and a binary cross-entropy loss function that compares those probabilities with true labels for error computation and to promote convergence during training. However, in this work, one of the loss functions employed during training was *BCEWithLogits*, which already combines both the sigmoid activation and the binary cross entropy loss into a **single operation**, ensuring numerical stability by doing the logits directly within the loss function. This avoids extreme probability values (very close to 0 or 1) during the binary cross-entropy calculation and helps prevent vanishing gradients, caused by unstable logits.

4.3.3 Training protocol

The methodology of this work was split into two phases according to the utilised datasets. More details about the methods applied to each dataset are provided in the following subsections. Moreover, a deterministic approach was ensured by using a fixed random seed (42) in each training process, so the randomness from splitting the dataset into subsets was reproducible.

Despite the remote workstation being equipped with 4 GPUs, 128 cores, and 256 GB of RAM, training and testing were conducted on a single GPU at a time using PyTorch version 2.5.1 (Python version 3.13.1).

4.3.3.1 Development set: pelvis studies from SynthRAD2023

Model training and validation were performed using a stratified 5-fold CV technique. This sampling technique randomly divided the training dataset into 5 folds (162 patients). In each iteration, 4 folds (80% - 129 patients) of the training subset were used for training, while the remaining 1 fold (20% - 33 patients) was retained as the validation subset for testing the model. The CV process was repeated 5 times, with no overlap between training and validation sets in each fold. This method, which stratifies the samples by centres, ensures that the original distribution of each class (centres A and C) is maintained throughout each fold and iteration.

It should be noted that the training folds comprised 3D images from which a single 2D slice with a resolution of 256×256 was randomly selected at each iteration (a total of 12900 iterations). This process simplifies the training process and reduces memory usage, allowing for more experiments in less time, since training with full 3D volumes at each iteration would be more time-consuming. Conversely, each batch of the validation fold treated 3D images as a collection of all 2D slices (each slice with 256×256 pixels), with slices subsampled at a predefined rate. The advantage of this approach is that it reduces memory usage while allowing almost all slices of the volumetric images to be evaluated by the model.

For validation inference, the generator used the *SlidingWindowInferer* method, which divided the input MRI into non-overlapping windows of size 256×256. The generator model independently evaluated the patches, sliding the entire image to compute predictions. Each non-overlapping window prediction was assigned the same weight, contributing equally to the final output.

To avoid the presence of tiling artefacts on the images created, the Gaussian mode was employed in the *SlidingWindowInferer* method for literature comparison, during testing. In this case, the default parameters of the method were used, evaluating overlapping windows with 25% overlap with a sigma value of 12.5%., which is the width of the Gaussian weighting function compared to its edges, and computing predictions of the various windows. Predictions in the overlapping regions were given less weight to mitigate boundary artefacts, whereas predictions near the window centre were assigned higher weights to follow a Gaussian distribution. As a result, this method smoothed transitions and reduced edge artefacts.

Hyperparameter tuning

Both the AE and cGAN models were subjected to a cascade-type process, in which several parameter combinations were tested progressively, as summarised in Table 4.4 and Table 4.5. The aim was to achieve a combination of HP that produced the best metrics, and therefore the best synthetic-CT images. Progression along the cascade was done when the best value or type was found, based on four image metrics (PSNR, MAE, SSIM, and MS-SSIM). These metrics were calculated as the mean of the 5 models' predictions, each validated on its respective subset. Unlike the CV process, this evaluation did not use the subsample rate of slices, relying on the *SlidingWindowInferer* method (described above - validation inference). If **all** the new metrics outperformed the ones from the previous stage, the respective HP combination was selected.

When this was not the case, an average ranking was computed between the HP combinations for each metric, and the best-ranked combination was chosen.

Initially, the control of the overall learning process was conducted with the following **baseline** set of HP for AE:

- Subsampling rate of slices for the validation fold images of 2; Adam optimizer was employed with the default parameters, specifically $\beta_1 = 0.9$ and $\beta_2 = 0.999$; trained for 100 epochs; MAE loss function; U-Net comprising 5 layers, with batch normalisation at the end of each layer; batch size for training was set to 32; batch size for validation was set to 1.

Table 4.4: HP explored in the AE model. Abbreviations: Batch Normalisation (BN), Instance Normalisation (IN), Rectified linear unit (ReLU), Leaky rectified linear unit (LeakyReLU).

Hyperparameter	Values/Types explored
Learning rate	[0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.0006, 0.001, 0.002, 0.005, 0.01]
Architecture - Number of layers	[6, 7, 8]
Architecture - Residual units per layer	[0, 2]
Architecture - Normalisation type	[BN, IN]
Training batch size	[32, 8, 1]
Architecture - Activation function type	[ReLU, LeakyReLU]
Epochs	[200, 800]

Bellow are the **baseline** group of HPs used as a starting point for cGAN tuning:

- Subsampling rate of slices for the training images of 7; Adam optimizer was employed with $\beta_1 = 0.5$ and $\beta_2 = 0.999$; trained for 200 epochs; $\lambda_{MAE} = 100$ (for the MAE loss function); $\lambda_{BCEwithLogists} = 1$ (for the BCEWithLogists loss function); $\lambda_{PL} = 0$ (for the PL function); U-Net comprising 5 layers, in which the activation function was ReLU, the kernel size was set to 3, and with the batch normalisation at the end of each layer; discriminator architecture with LeakyReLU activation function (slope=0.2), kernel size equal to 4, without dropout, and batch normalisation in each block (2D convolution Layer - Batch normalisation layer - LeakyReLU activation layer), except the first one; batch size for training was set to 32; batch size for validation was set to 1.

Table 4.5: Hyperparameters investigated for the cGAN model. Abbreviations: Generator (G), Discriminator (D), Batch Normalisation (BN), Instance Normalisation (IN), Rectified linear unit (ReLU), Leaky rectified linear unit (LeakyReLU).

Hyperparameter	Values/Types explored
Learning rate (G=D)	[0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.0006, 0.001, 0.002, 0.005, 0.01]
Architecture (G) - Number of layers	[6, 7, 8]
Architecture (G) - Residual units per layer	[0, 2]
Architecture (G and D) - Normalisation type	[BN(G)+BN(D), IN(G)+IN(D)]
Training batch size	[32, 8, 1]
Architecture (D) - Kernel size	[3, 4]
Architecture (G) - Activation function	[ReLU, LeakyReLU]
Learning rate (D)	[0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0006, 0.001, 0.002, 0.005, 0.01]
Loss function - Perceptual loss weight (λ)	[0, 0.5, 1, 10]
Improved training technique - One-Sided label smoothing (D)	[D(real)=1.0 and D(fake)=0.0 D(real)=0.9 and D(fake)=0.0]
Architecture (D) - Dropout rate	[0, 0.25]
Epochs	[200, 800]

An improved training technique known as one-sided label smoothing (OSLS) was included in the HP search [62]. This method assists in mitigating the discriminator's overconfidence by setting the target labels for real images to 0.9 rather than 1.0. This encourages more stable training, as recommended by Salimans et al. [62]. During the HP tuning, the perceptual loss selected used features from a pretrained deep neural network (ResNet) trained on the RadImageNet dataset, due to its efficient transfer learning in artificial intelligence applications in radiologic imaging. The dataset includes 1.35 million annotated medical images in 131872 patients who underwent CT, MRI, and for musculoskeletal, neurologic, oncologic, gastrointestinal, endocrine, abdominal, and pulmonary pathologic conditions [63]. Thus, high-level features are extracted from CT and sCT images using a pretrained neural network to evaluate the perceptual differences that the network has learned to detect.

Regarding the order of the HP tuning, which depended on the model, these followed a logical sequence, adapted according to the observed training results.

Ensemble of models

In order to understand whether the predictions of the five individual models, which were derived from the stratified 5-fold CV with the final HP configuration, outperformed those of the ensemble of models, all models were tested on unseen images from 18 patients (hold-out test set). More specifically, the ensemble of models approach consisted of computing the mean predictions across the five individual models for each image. The evaluation employed to compare the ensemble against the predictions of individual models, was carried out with four image quality metrics (PSNR, MAE, SSIM, MS-SSIM), alongside a paired sample t-test.

Performance comparison: AE vs. cGAN

The predictions of the ensemble of AE and cGAN models were compared to determine which architecture better generated realistic sCT images for a small mono-region dataset.

4.3.3.2 Multi-region set: SynthRAD2025

Given the best HP formula obtained with the SynthRAD2023 dataset, a new multi-region, multi-centre and larger dataset (SynthRAD2025) was used to evaluate two distinct approaches: a centre-stratified approach, where training images were stratified by centre and a region-based approach, where the models were trained either on individual regions (region-specific) or on multiple regions simultaneously (multi-region). Table 4.6 illustrates each train/test combination for the region-based approach. Concerning the training process, this time the number of epochs was increased to 1000 to account for a more diverse dataset, which included multiple regions and more images.

In relation to the two approaches, the CV training and validation sets were cropped to dimensions of 256×256 pixels, where for training folds, a random 2D slice was selected per iteration (46250 in total for centre-stratified approach and 46375 in total for the multi-region), and for validation folds, as opposed to the SynthRAD2023 dataset, no subsampling rate was applied. This contributed to a more robust validation process during training, since the evaluation was performed using the entire set of 2D slices from each patient. The inference strategy followed the same approach used for the SynthRAD2023 dataset.

Centre-stratified approach

Similarly to the SynthRAD2023, a 5-fold CV stratification on the centres was implemented. The whole training dataset was partitioned into 5 folds (463 patients), with 4 folds (80% - 370/371 patients, depending on the fold division) for training and 1 fold (20% - 93/92 patients, depending on the fold division) to validate the model, at each iteration. This process was repeated five times, in which a totally different validation set was used to evaluate the model for both the AE and cGAN models.

Train/test combinations for the centre-stratified approach, data from centres A, B, and C were combined for both training and validation. Testing was performed on the remaining merged data from centres A, B, and C, referred to as **internal** testing, while the evaluation using unseen data from centre D was referred to as **external** testing.

To investigate the generalisability of the model on new inputs, internal testing was performed on 50 patients, whose images were from all anatomical regions (abdominal, thoracic, and head and neck) and centres A, B, and C, and external testing on 65 patients who acquired CT and MRI scans of the head and neck region in centre D.

Performance comparison: AE vs. cGAN

The ensemble predictions of 50 patients from the internal test (from A, B and C centres) were used to estimate and compare AE and cGAN performances.

Region-based approach

In this study, two types of training approaches were used with the AE model: i) models trained on multiple regions, and ii) models trained on individual regions, one at a time. For **models trained in all regions**, the 5-fold CV was stratified by regions (abdominal, thoracic, and head and neck), totalling approximately 464 patients (5 folds). The division in folds included 4 training folds with 371 patients and 1 validation fold with 93 patients. Therefore, the 4 training folds/1 validation fold split was carried out in the following way:

- Model trained on AB region: 127 training patients (total of 15875 slices across 1000 epochs)/32 validation patients
- Model trained on TH region: 131 training patients (total of 16375 slices across 1000 epochs)/33 validation patients
- Model trained on HN region: 113 training patients (total of 14125 slices across 1000 epochs)/28 validation patients

Unlike the multi-region approach, the training process for the **models trained in individual regions** did not use stratification (5-fold CV). The distribution of the images per region in the region-specific study matched the previously defined multi-region training/validation split. In this case, the goal was to analyse the performance of region-specific models compared to the multi-region model.

Table 4.6: Train/Test combinations for the region-specific approach. The datasets with HN (head and neck - centres A, B, and C), AB (abdominal), and TH (thoracic) images were trained and evaluated independently and aggregated. Data from external centre D was solely used for testing, designated as HN (external).

Train / Test	AB	HN	HN (external)	TH	AB+HN+TH
AB	✓				
HN		✓	✓		
TH				✓	
AB+HN+TH	✓	✓	✓	✓	✓

For testing, the same 49 patients were used in both models. The internal evaluation was conducted on all 49 patients from the three regions using the multi-region model. In the region-specific models, the internal tests used the 49 patients separately: 16 patients from the AB region (centre A: 6; centre B: 9; centre C: 1), 18 patients from the TH region (centre A: 9; centre B: 9), and 15 patients from the HN region (centre A: 9, centre C: 6). External evaluation was performed on 65 cases from the HN region (centre D), using both the region-specific and multi-region models.

4.4 Performance assessment

The synthetic-CT images were evaluated in three main cohorts: imaging similarity (applied to both SynthRAD2023 and SynthRAD2025 images), geometric accuracy (applied only to a subset of SynthRAD2025 images), and dose calculation performance (also applied only to a subset of SynthRAD2025 images).

4.4.1 Imaging similarity

To evaluate the quality of the sCT images, MAE, PSNR, SSIM, and MS-SSIM image quality metrics were computed by the expressions previously described in Table 3.4 of the State of the art chapter. These metrics were applied to all 2D **masked** slices of real CT and pseudo CT volumes (each volume with shape [D, 1, H, W]), rather than to the full 3D volume scans ([B, 1, H, W, D], where B is the batch size). Although 3D volume evaluation could be considered, to ensure consistency with the training on randomly selected 2D slices, the evaluation of the model was done by averaging metrics computed across all 2D slices ([D, 1, H, W]), which offers a fairer assessment of the model's performance. This framework has the advantage of capturing local variations across slices, due to per-slice variability. On the other hand, if the metrics were computed on single slices, they would not consider their spatial relationship, since they belong to the same patient. Except for the MAE, all other metrics were calculated on normalised images (not in HU).

In both SSIM and MS-SSIM, the k_1, and k_2 constants were set at 0.01 and 0.03 for $C_1 = (k_1 Q)^2$ and $C_2 = (k_2 Q)^2$, respectively. In more detail, SSIM metric measures the structural similarity between two images in terms of luminance, contrast and structure comparisons, by using a sliding window approach, where the window moves one pixel at a time across the entire image space. At each position, the SSIM index is computed within that local window. To avoid blocking artifacts caused by an NxN square window, a smooth windowing approach is employed for local statistics. In practice, this metric aims to mimic the human visual system, whose subjective evaluation of an image changes depending on the sampling density of the image signal, the distance from the image plane to the observer, and the perceptual capability of the observer's visual system [53].

To capture image details at different resolutions, the MS-SSIM was introduced, as an extension of SSIM. Unlike SSIM, MS-SSIM metric computes the similarity measure at different

resolution levels, denoted by M. For that, the system iteratively creates lower-resolution versions of the image by applying a low-pass filter (blurs the image very slightly) followed by downsampling the filtered images by a factor of 2 (reduces the image dimension by half), in each of the M-1 iterations. At each scale, contrast and structure comparisons are calculated, while the luminance comparison is computed only at the highest scale (M). In this work, the number of resolution levels was set to 5 ($M = 5$) [53].

4.4.2 Geometric consistency

To mimic the SynthRAD2025 challenge evaluation workflow, the CT and sCT images generated from **the model trained in all regions** from SynthRAD2025 were automatically segmented into specific anatomical structures (Table 4.7) using the TotalSegmentator tool [64]. This tool was designed to segment 104 anatomical structures (27 organs, 59 bones, 10 muscles, and 8 vessels) in any CT image, and is based upon an nnU-Net framework [65], which is a state-of-the-art model. Trained in a dataset comprising 1204 CT scans from patients of different ages, abnormalities, acquired on different scanners, body parts, sequences, and sites, the TotalSegmentator tool is robust and works well in most images [64].

Table 4.7: Regions of interest segmented from CT and sCT with TotalSegmentator tool. All segmented classes were saved in a single NifTi file from the task "total" in TotalSegmentator, for each patient. For the HN region, the number of thoracic vertebrae varied across patients from centre A, while patients from centre C had only cervical vertebrae included in their scans.

Region	Anatomical segments
AB/TH	Left and right kidneys, liver, stomach, lung lobes, vertebrae (1 sacral, 5 lumbar, 12 thoracic and 7 cervical), heart, spinal cord, ribs (12 on the left side and 12 on the right side), sternum
HN	Oesophagus, trachea, thyroid, vertebrae (7 cervical and 12 thoracic), spinal cord, brain, skull

The segments were compared based on two metrics, the mDICE and the HD95, with the following expressions:

- **multiclass-DICE coefficient (mDICE)**: measures the similarity between the segmentation of each class (type of segment) in CT and sCT.

$$mDICE = \frac{1}{N} \sum_{i} \frac{2 \left| \text{Seg}_{CT,i} \cap \text{Seg}_{sCT,i} \right|}{\left| \text{Seg}_{CT,i} \right| + \left| \text{Seg}_{sCT,i} \right|} \tag{4.1}$$

- **95th pertencile Hausdorff distance (HD95)**: quantifies the dissimilarity between segment boundaries, meaning that it measures the distance from each boundary point in

Seg_{CT} to the nearest point on the boundary of Seg_{sCT}, reporting the 95th percentile of these distances.

$$HD(\mathrm{Seg}_{X,i}, \mathrm{Seg}_{Y,i}) = \max \left\{ \sup_{x \in \mathrm{Seg}_{X,i}} d(x, \mathrm{Seg}_{Y,i}), \sup_{y \in \mathrm{Seg}_{Y,i}} d(\mathrm{Seg}_{X,i}, y) \right\} \tag{4.2}$$

4.4.3 Dose comparison

Treatment planning

Given that the challenge authors did not provide any treatment plans and for the sake of comparison, the same treatment plan parameters were established for all patients. Dose calculations were performed using matRad, an open-source treatment planning system based on MATLAB for research and educational purposes [33]. For this work, only IMPT plans were implemented with planning parameters based on the SynthRAD2025 challenge guidelines. Although IMPT is not the standard treatment type for all cancers selected for dose analysis, its deposition of maximal energy at a specific tissue depth decreases the dose delivered to OARs, offering an advantage over IMRT.

Firstly, the CT and sCT scans, which were generated by **the model trained in all regions** of SynthRAD2025, were imported to matRad and then, their HU values were converted to the corresponding relative stopping power in water, crucial for dose calculations in proton therapy, with matRad's default calibration curve. This calibration curve is based on the HLUT, depicted in Figure 2.7 from the Background Concepts chapter.

Secondly, a setup with three proton beams was defined (Table 4.8). The beam geometry included gantry angles of 0^{o}, 120^{o} and 240^{o}, couch angles fixed at 0^{o} for all beams, and isocenters located at the centre of mass of the GTV for each patient. The number of fractions depended on the region: 30 fractions for the AB and HN regions and 35 fractions for TH region [66]. Regarding the radiation delivery system, this was set to a generic machine, as specified by the SynthRAD2025 challenge, which also defined the dose prescriptions per fraction of 2 Gy, the number of fractions, the number of beams, and objective functions and constraints [66]. Other treatment planning parameters, such as the optimisation quantity, proton-specific dose calculation algorithm, and dose grid resolution, were adopted from a proton therapy matRad example, as the assigned penalties of 1000 for the GTV. Once the beam geometry and all parameters were defined, each pencil beam dose contribution was calculated and saved in dose influence matrices.

Table 4.8: Treatment planning parameters used in this work.

Parameters	Setting
Number of fractions	30 (AB/HN) e 35 (TH)
Radiation mode	Protons
Machine model	Generic
Biological model	Constant relative biological effectiveness (RBE)
Scenario type	Nominal
Bixel width	5mm
Gantry angles	0^{o}, 120^{o}, 240^{o}
Couch angles	0^{o}, 0^{o}, 0^{o}
Number of beams	3
Dose calculation algorithm	Hong pencil beam
Dose grid resolution	$3mm \times 3mm \times 3mm$
Optimization quantity	RBE-weighted dose

Next, fluence optimisation was performed only on sCT based on clinical objectives and constraints (Table 4.9) for the GTV, and OARs, to find the optimal set of beamlet weights that produced the best dose distribution. In the dose calculation applied to GTV structures, a mean squared deviation objective function was used, in which the difference between the delivered dose and prescribed dose is squared, while the OARs were calculated based on hard constraints to limit dose exposure strictly. The optimisation only happened on the sCT to simulate a real MR-only radiotherapy workflow, in which only the synthetic-CTs is available to dose calculations. Lastly, the optimal beamlet weights were applied to the corresponding CT to calculate its dose distribution, without further optimisation. This strategy enabled the analysis of plan optimisation on the sCT, since both images shared the same treatment plan. Therefore, any dose differences between sCT and CT were exclusively caused by differences in HU.

Moreover, given the absence of treatment plans, the desired structures were segmented from CT images using the TotalSegmentator tool [64], since segmenting directly from sCT could introduce inaccuracies. These segmentations were used as GTV and OARs in both the CT and sCT dose calculations to ensure a consistent reference and a fairer comparison.

The cancer sites with the highest incidence per anatomical region were selected as GTV, except for the HN region. According to the global cancer statistics database of 2022 by the International Agency for Research on Cancer (IARC), which is part of the World Health Organisation (WHO) [67], the most incident cancers within each anatomical region are the following:

- Abdominal region: Stomach cancer, ranked 5th among all cancer types, was used as GTV.

- Thoracic region: Lung cancer, the most incident cancer type among all (ranked 1st). The GTV chosen was the right upper lobe of the lung, since it showed the most pronounced increase in incidence among the lung lobes [68].
- Head and neck region: Thyroid gland cancer, ranked 7th in all types of cancers. However, this structure segment was not available in patients from centre C. As a result, oral cavity cancers (including the lip), ranked 16th overall, were used instead. In this case, the GTV consisted of three structures within the oral cavity: the tongue, the hard palate and soft palate, which were selected based on their availability in the TotalSegmentator tool [64].

Dosimetric evaluation

To analyse and examine the plans based on CT and sCT generated from **the model trained in all regions**, three dose distribution metrics were employed:

1. **Mean absolute dose difference ($MAE_{\textbf{target dose}}$)**: determines the average absolute differences in dose values between both CT and sCT within the region that receives at least 90% of the prescribed dose, relatively to the prescribed dose (2 Gy per fraction). The variable n represents the number of voxels within this region and i the voxel index.

$$\mathrm{MAE}_{\text{target dose}} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{D_{\geq 90\%,\mathrm{CT},i} - D_{\geq 90\%,\mathrm{sCT},i}}{D_{\text{prescribed}}}\right| \tag{4.3}$$

2. **Dose-volume histogram metric ($DVH_{\textbf{metric}}$)**: considers the absolute difference between a specific DVH parameter in CT and sCT, relative to the corresponding parameter in CT. The four considered parameters are: $D98\%_{target}$ (dose that the target volume received in at least 98% of its volume), $V95\%_{target}$ (target volume receiving at least 95% of the prescribed dose), $D2\%_{OAR}$ (dose that 2% of the volume at a specific OAR received), and $Dmean_{OAR}$ (mean dose that was received by a specific OAR). Regarding the OARs parameters, these were the mean absolute difference, with n_{OARs} corresponding to the number of OAR, which is 3. These 3 OARs were selected by computing the mean of D5% and $Dmean$ for each organ and choosing those three with the highest mean values. Consequently, these correspond to the OARs that received the highest doses.

$$\begin{aligned} DVH_{\text{metric}} = &\left|\frac{D98\%_{\text{target,CT}} - D98\%_{\text{target,sCT}}}{D98\%_{\text{target,sCT}}}\right| + \left|\frac{V95\%_{\text{target,CT}} - V95\%_{\text{target,sCT}}}{V95\%_{\text{target,sCT}}}\right| \\ &+ \frac{1}{n_{\mathrm{OARs}}}\sum_{\mathrm{OARs}}\left|\frac{D2\%_{\mathrm{OAR,CT}} - D2\%_{\mathrm{OAR,sCT}}}{D2\%_{\mathrm{OAR,sCT}}}\right| \\ &+ \frac{1}{n_{\mathrm{OARs}}}\sum_{\mathrm{OARs}}\left|\frac{Dmean_{\mathrm{OAR,CT}} - Dmean_{\mathrm{OAR,sCT}}}{Dmean_{\mathrm{OAR,sCT}}}\right| \end{aligned} \tag{4.4}$$

3. **Gamma index ($\gamma(r_{sCT})$)**: quantitatively evaluates how similar two 3D dose distributions are by considering the following criteria: the distance-to-agreement, Δd_M, which measures the distance between a point on the measured dose distribution and the nearest point on the reference dose distribution, and the dose-difference, ΔD_M, which measures how the

Table 4.9: Planning objectives (GTV) and dose constraints (OARs) used in matRad for each region. Abbreviations: RUL - right upper lobe, RML - right middle lobe, RLL - right lower lobe, LUL - left upper lobe, LLL - left lower lobe, R - right, L - left, S - superior, M - middle, I - inferior.

HN: 30x2Gy (Oral cavity)		**TH: 35x2Gy (RUL)**		**AB: 30x2Gy (Stomach)**	
Structure	**Constraint**	**Structure**	**Constraint**	**Structure**	**Constraint**
Spinal cord	$D_{0.03cc} < 48Gy$	Spinal cord	$D_{0.03cc} < 50.5Gy$	Spinal cord	$D_{0.03cc} < 50.5Gy$
Brain stem	$D_{0.03cc} < 54Gy$	Oesophagus	$D_{0.03cc} < 73.5Gy$ $V_{60Gy} < 15.3\%$ $D_{mean} < 30.6Gy$	Liver	$D_{mean} < 30Gy$
Optical nerve (R/L)	$D_{0.03cc} < 54Gy$	RML, RLL LUL,LLL	$V_{60Gy} < 33\%$ $V_{5Gy} < 33\%$ $D_{mean} < 18Gy$	Kidney (R/L)	$D_{mean} < 18Gy$
Parotid glands (R/L)	$D_{mean} < 25Gy$			Urinary bladder	$D_{0.03cc} < 65Gy$
Larynx	$D_{mean} < 40Gy$				
Submandibular glands (R/L)	$D_{mean} < 39Gy$				
Pharynx constrictor (S/M/I)	$D_{mean} < 50Gy$				
Common carotid artery (R/L)	$D_{0.03cc} < 40Gy$				
Eye lens (R/L)	$D_{0.03cc} < 5Gy$				
Masseter (R/L)	$D_{mean} < 35Gy$				
Oesophagus	$D_{mean} < 30Gy$				
Thyroid	$D_{mean} < 40Gy$				

measured and reference dose differ at those points. In this case, the dose distribution in the real CT was the evaluated/measured distribution, and the dose distribution in the synthetic-CT treatment plan was used as the reference/planned dose distribution. The resulting gamma index computed for each voxel in the 3D dose distributions is defined as:

$$\gamma(\mathbf{r}_{\mathrm{CT}}) = \min\left\{\Gamma(\mathbf{r}_{\mathrm{sCT}}, \mathbf{r}_{\mathrm{CT}})\right\} \forall \left\{\mathbf{r}_{\mathrm{CT}}\right\}, \tag{4.5}$$

where Γ is defined as:

$$\Gamma(\mathbf{r}_{\mathrm{sCT}}, \mathbf{r}_{\mathrm{CT}}) = \sqrt{\left(\frac{|\,\mathbf{r}_{\mathrm{CT}} - \mathbf{r}_{\mathrm{sCT}}|}{\Delta d_M}\right)^2 + \left(\frac{D_{\mathrm{CT}}(\mathbf{r}_{\mathrm{CT}}) - D_{\mathrm{sCT}}(\mathbf{r}_{\mathrm{sCT}})}{\Delta D_M}\right)^2} \tag{4.6}$$

Thus, the gamma index is the minimum gamma value calculated over all points from the measured distribution. The Figure 4.8 shows the fundamental principle of gamma analysis. In more detail, the gamma index calculation involves defining a 2D space around one axis being the dose difference (ΔD_M) and the other being the distance-to-agreement (Δd_M), as well as drawing a circle with a radius of 1 around each reference point in that space, or a sphere for 3D distributions. Then, the Euclidean distance between each measured dose distribution point and the reference point, normalised to the passing criteria, is computed, yielding multiple gamma index values across the evaluated region. Among the calculated values, the **minimum gamma value** is the most important metric to assess clinical acceptability.

Other parameters, such as the passing rate, are computed to assess the outcome of the gamma analysis in 3D. A threshold is used to determine whether the entire dose distribution passes or fails the gamma analysis, and is reported as a percentage. If the percentage of voxels within the evaluated region satisfies $\gamma(\mathbf{r}) \leq 1$, and it is greater than a threshold equal to 10% of the $D_{prescribed}$ (2 Gy), the treatment plan is clinically accepted. In contrast, if $\gamma(\mathbf{r}) > 1$, the plan is rejected. In this work, the criteria used was $\Delta d_M = 2mm$ and $\Delta D_M = 2\%$, commonly referred to as 2%/2mm criterion, similar to the SynthRAD2023 and SynthRAD2025 challenge guidelines [55] [66].

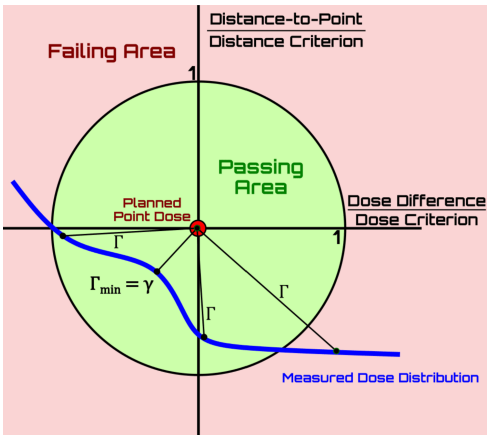


Figure 4.8: Gamma index representation in the 2D space. Image from [69].

4.4.4 Statistical comparisons

Statistically significant differences were investigated using the paired sample t-test to compare the means of two related groups. Since the null hypothesis (H_0) states that the difference in the means of two related groups is zero, for $p_{value} < 0.05$, the H_0 is rejected, indicating that the paired means are significantly different. This test was applied to: i) compare the performance between AE and cGAN models on the two datasets (SynthRAD2023 and SynthRAD2025); ii) compare the mean metrics of a single model versus an ensemble of models with violin plots; and iii) compare models trained in individual regions with the ones trained in all anatomical regions when validated on images from a specific region. Additionally, violin plots offer an advantage over the box plot by displaying not only all the same statistics (median, quartiles, and outliers), but also the entire distribution of the data.

Chapter 5

Results

This chapter describes the results from model training, evaluation, ensemble approach, and performance comparisons using both datasets (sections 5.1 and 5.2). Section 5.2 also includes the comparison between one-region and multi-region models, as well as the validation results of the multi-region model in terms of 3D imaging quality, geometric consistency and dose metrics.

5.1 SynthRAD2023

5.1.1 Hyperparameter search

After training several models (grouped by different sets of HP) and evaluating them using a stratified 5-fold CV, the model with the best performance was selected and tested on the validation set. For the AE and cGAN, the best-performing model was saved when the weights of the epoch with the lowest average MAE loss (mean loss computed over batches) were reached. The assessment of the best model performance was based on using image quality metrics. Using the same set of data (validation set) for model selection and validation can lead to biased results; as such and as noted in the Methods section, we made use of a separate hold-out test set for metric calculation of the best final models. Figure 5.1 and Table 5.1 illustrate the HP search process through each optimization stage and its percentage of change relative to the baseline for the **AE model**, with the graph in Figure 5.1 demonstrating a generally monotonic trend until the optimization of the normalization type, followed by a rapid increase from there on. Since the stages are cascaded, the final stage refers to the optimal HP configuration, as shown in Table 5.2. The validation metrics obtained in that stage are detailed in Table 5.3 for AE. As shown in Table 5.1, at each optimization step, PSNR increased. In contrast, MAE consistently decreased, improving image quality. Even though the SSIM percentage of change from baseline also increased in every stage, this trend was not significantly reflected in the mean metric values, specifically in the number of layers and normalization type tuning. For the MS-SSIM, the calibration of the number of layers slightly diminished the difference in percentage from the baseline, which was not mirrored in the mean MS-SSIM value. The SSIM, a meaningful increase in the mean metric value was only observed when refining the training batch size and the number of epochs.

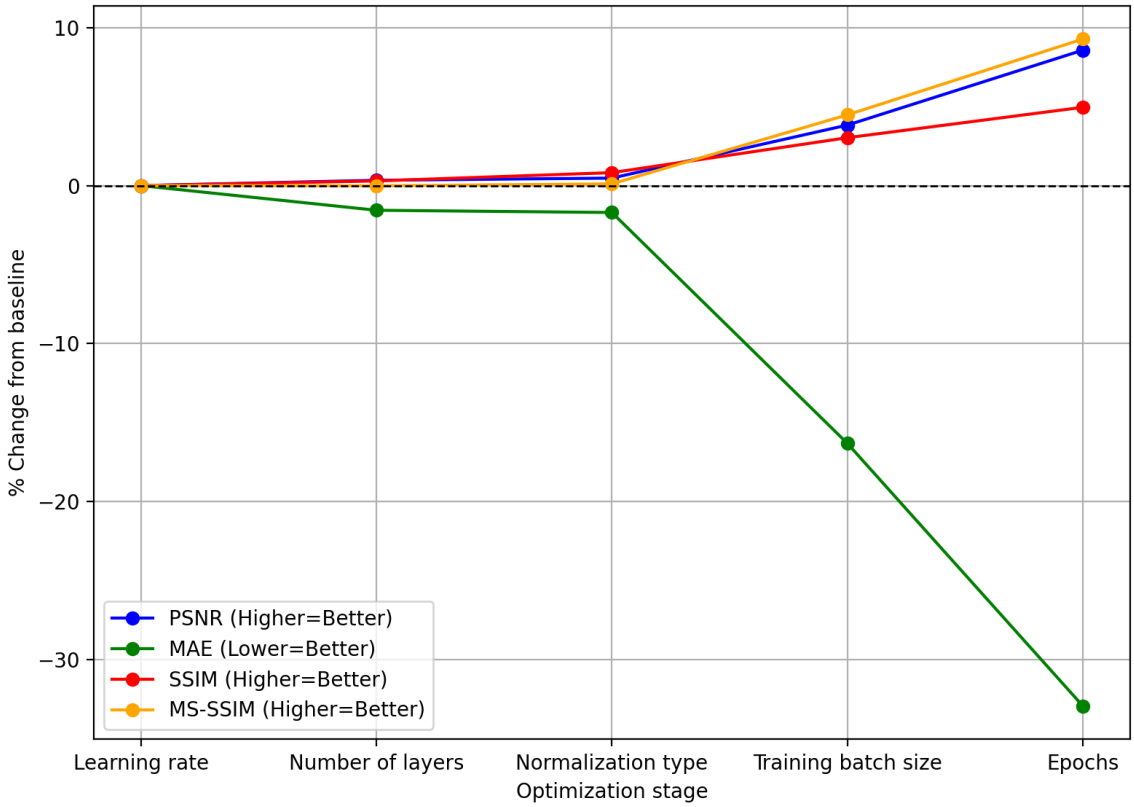


Figure 5.1: Progression of the percentage change from baseline in validation metrics performance (PSNR, MAE, SSIM, MS-SSIM) across hyperparameter optimization stages for the AE model. The difference in percentage was calculated using the average of 5-fold cross-validation means.

Table 5.1: Validation metrics performance and percentage of the difference from baseline across optimization stages for the AE model. The best metrics are highlighted in bold.

Optimization stage	PSNR		MAE		SSIM		MS-SSIM	
	(dB,↑)	%Δ	(HU,↓)	%Δ	(-,↑)	%Δ	(-,↑)	%Δ
Learning rate (Baseline)	26.12	0.00%	103.72	0.00%	0.88	0.00%	0.81	0.00%
Architecture - Number of layers	26.21	0.34%	102.10	-1.56%	0.88	0.30%	0.81	-0.01%
Architecture - Normalization type	26.24	0.48%	101.95	-1.70%	0.88	0.82%	0.81	0.11%
Training batch size	27.12	3.83%	86.79	-16.33%	0.90	3.03%	0.85	4.48%
Epochs	**28.36**	8.58%	**69.54**	-32.95%	**0.92**	4.96%	**0.89**	9.28%

Table 5.2: Hyperparameters identified as those achieving best imaging metrics in the AE model. Abbreviations: Instance Normalization (IN), rectified linear unit (ReLU).

Hyperparameter	Best value/type
Learning rate	0.001
Architecture - Number of layers	8
Architecture - Residual units per layer	0
Architecture - Normalization type	IN
Training batch size	8
Architecture - Activation function type	ReLU
Epochs	800

Table 5.3: Stratified 5-fold CV image metrics results utilizing the AE model on the SynthRAD2023 dataset. Values across all folds for SSIM and MS-SSIM metrics are consistent, while in PSNR and MAE, the stratification in fold 5 produced the best metrics.

	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
Fold 1	28.32 ± 2.00	73.54 ± 15.28	0.92 ± 0.02	0.89 ± 0.03
Fold 2	28.30 ± 1.97	70.24 ± 14.35	0.92 ± 0.01	0.89 ± 0.03
Fold 3	28.34 ± 1.99	70.36 ± 13.57	0.92 ± 0.02	0.89 ± 0.03
Fold 4	28.30 ± 1.94	67.20 ± 10.96	0.92 ± 0.02	0.89 ± 0.03
Fold 5	28.54 ± 1.61	66.36 ± 9.70	0.92 ± 0.01	0.89 ± 0.02
Mean	**28.36 ± 1.91**	**69.54 ± 13.22**	**0.92 ± 0.01**	**0.89 ± 0.03**

Figure 5.2 and Table 5.4 present the hyperparameter process for the **cGAN model** through each optimization stage and its percentage of change from baseline. Following the same cascade architecture as the AE model, the final stage represents the optimal hyperparameter arrangement, with its final set of HP in Table 5.5, and validation metrics provided in Table 5.6. Overall, Figure 5.2 presents a steady improvement across all metrics, with the following exceptions: the discriminator learning rate, perceptual loss weight, and one-sided label smoothing technique that culminates in a plateau.

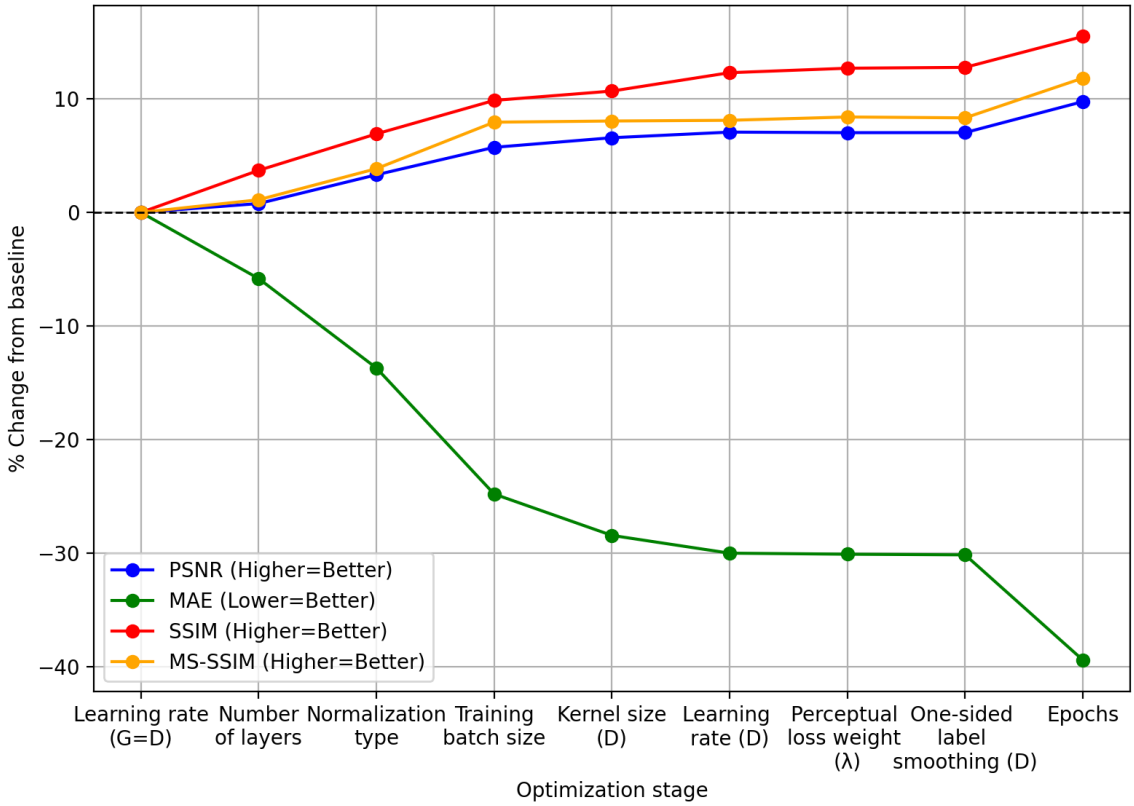


Figure 5.2: Progression of the percentage change from baseline in validation metrics performance (PSNR, MAE, SSIM, MS-SSIM) across hyperparameter optimization stages for the cGAN model. The difference in percentage was calculated using the average of 5-fold cross-validation means.

As depicted in Table 5.4, for the PSNR metric, calibrating the perceptual loss weight and introducing the one-sided label smoothing technique did not impact the results. Furthermore, optimizing the training batch size resulted in the largest percentage drop (-11.11%) among all stages, for the MAE values. From the training batch size until the one-sided label smoothing optimization stages, the metric values for MS-SSIM were not affected by the increment on the percentage of difference from baseline values. Almost identically, the SSIM values showed a marginal increase when the discriminator learning rate was optimized.

Table 5.4: Validation metrics performance and percentage of the difference from baseline across optimization stages for the cGAN model. The best metrics are highlighted in bold.

Optimization stage	PSNR		MAE		SSIM		MS-SSIM	
	(dB,↑)	%Δ	(HU,↓)	%Δ	(-,↑)	%Δ	(-,↑)	%Δ
Learning rate (G=D) (Baseline)	25.33	0.00%	130.62	0.00%	0.79	0.00%	0.79	0.00%
Architecture - Number of layers	25.53	0.78%	123.01	-5.83%	0.82	3.71%	0.80	1.12%
Architecture - Normalization type	26.17	3.32%	112.72	-13.70%	0.85	6.92%	0.82	3.86%
Training batch size	26.78	5.74%	98.22	-24.81%	0.87	9.87%	0.85	7.95%
Kernel size (D)	26.99	6.58%	93.46	-28.45%	0.87	10.69%	0.85	8.05%
Learning rate (D)	27.12	7.08%	91.42	-30.02%	0.89	12.31%	0.85	8.12%
Perceptual loss weight (λ)	27.11	7.03%	91.29	-30.11%	0.89	12.70%	0.85	8.41%
One-Sided label smoothing (D)	27.11	7.03%	91.22	-30.17%	0.89	12.78%	0.85	8.33%
Epochs	**27.80**	9.76%	**79.12**	-39.43%	**0.91**	15.50%	**0.88**	11.83%

Table 5.5: Hyperparameters identified as those achieving best imaging metrics in the cGAN model. Abbreviations: Generator (G), Discriminator (D), Instance Normalization (IN), rectified linear unit (ReLU).

Hyperparameter	Best value/type
Learning rate (G=D)	0.0005
Architecture (G) - Number of layers	7
Architecture (G) - Residual units per layer	0
Architecture (G and D) - Normalization type	IN(G)+IN(D)
Training batch size	8
Architecture (D) - Kernel size	3
Architecture (G) - Activation function	ReLU
Learning rate (D)	0.005
Loss function - Perceptual loss weight (λ)	1
Improved training technique - One-Sided label smoothing (D)	D(real)=0.9 and D(fake)=0.0
Architecture (D) - Dropout rate	0
Epochs	800

Table 5.6: Stratified 5-fold CV image metrics results utilizing the cGAN model on the SynthRAD2023 dataset. Except the metrics that remained nearly stable, such as SSIM and MS-SSIM, fold 2's training process achieved superior results in PSNR and MAE values.

	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
Fold 1	28.11 ± 2.00	77.70 ± 14.78	0.90 ± 0.02	0.87 ± 0.03
Fold 2	28.45 ± 2.01	69.16 ± 14.83	0.92 ± 0.01	0.89 ± 0.03
Fold 3	28.16 ± 2.01	75.83 ± 13.31	0.92 ± 0.02	0.89 ± 0.03
Fold 4	28.09 ± 1.92	72.45 ± 9.82	0.92 ± 0.02	0.89 ± 0.02
Fold 5	26.20 ± 2.34	100.44 ± 16.63	0.91 ± 0.01	0.87 ± 0.02
Mean	**27.80 ± 0.81**	**79.12 ± 11.06**	**0.91 ± 0.01**	**0.88 ± 0.01**

5.1.2 Ensemble of models

To determine whether testing the ensemble of models produced more realistic predictions than the individual models, this work compared the performances of the ensemble of five models with that of the individual models and evaluated them statistically with a paired sample t-test. As demonstrated in Figure 5.3 and Figure 5.4, the ensemble of models consistently outperformed the individual fold models across all metrics in both model architectures, with statistically significant differences using the paired sample t-test.

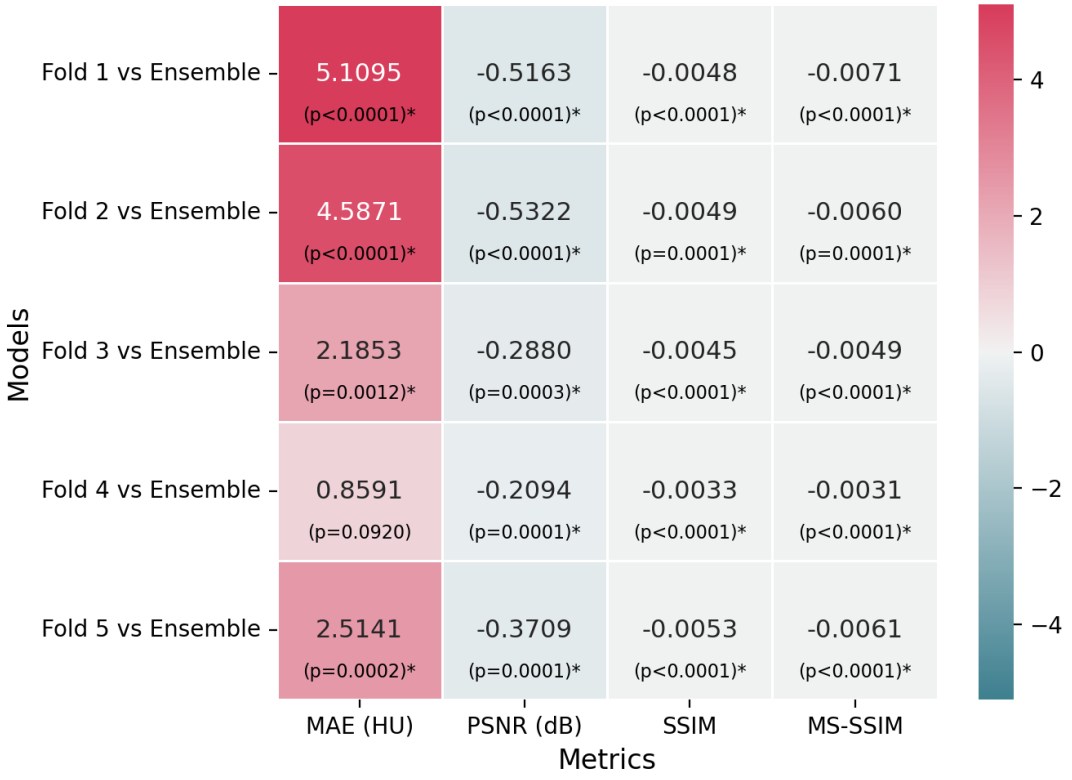


Figure 5.3: Performance comparison of metric differences (Fold-Ensemble) in the best performing AE model for SynthRAD2023 test set (18 patients). Heat map of mean metric differences between each fold and the ensemble of folds. Color intensity is divided into two colors: pink for positive differences and blue for negative differences. For the MAE metric, a pink tone means that the ensemble achieved lower (i.e., better) values. For the remaining metrics, the interpretation is the opposite: a blue tone indicates better performance of the ensemble. The statistically significant differences are denoted by asterisks ($p < 0.05$).

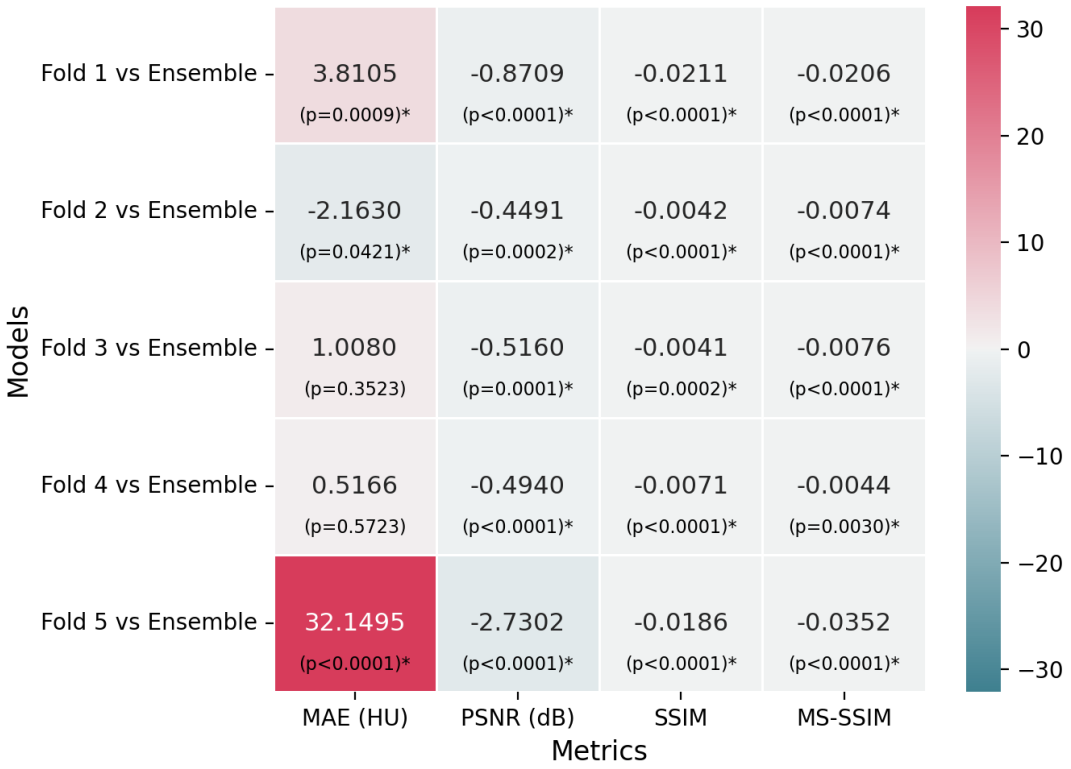


Figure 5.4: Performance comparison of metric differences (Fold-Ensemble) in he best performing cGAN model for SynthRAD2023 test set (18 patients). Heat map of mean metric differences between each fold and the ensemble of folds. Color intensity is divided into two colors: pink for positive differences and blue for negative differences. For the MAE metric, a pink tone means that the ensemble achieved lower (i.e., better) values. For the remaining metrics, the interpretation is the opposite: a blue tone indicates better performance of the ensemble. The statistically significant differences are denoted by asterisks ($p < 0.05$).

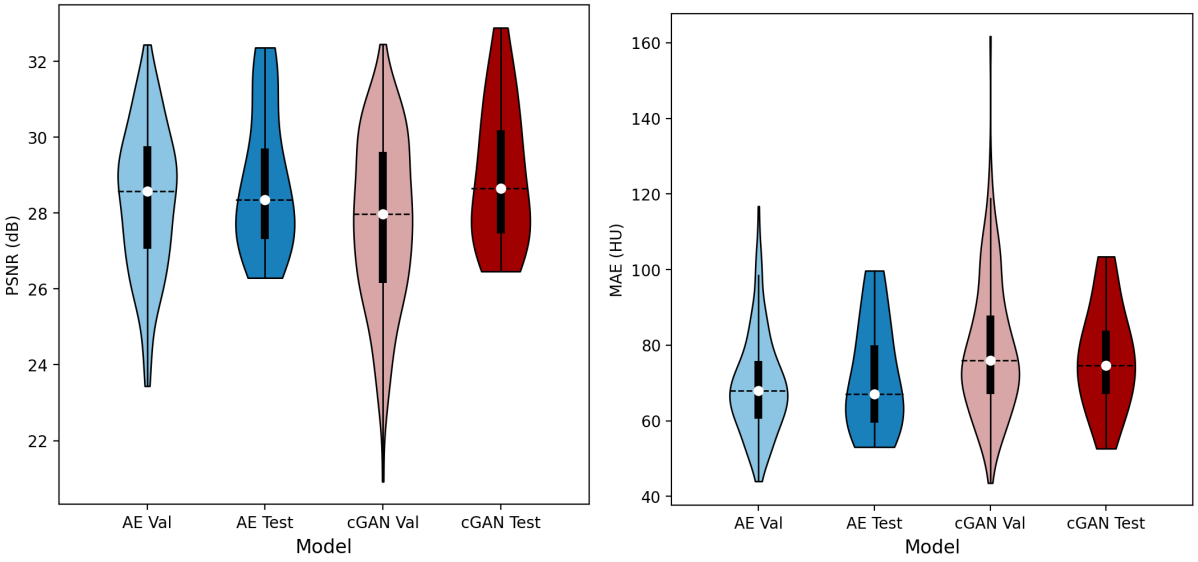
5.1.3 Performance comparison: AE vs. cGAN

Since the ensemble of models provided more realistic predictions, the comparison was performed between the ensembles of the best AE and cGAN models for testing. For validation, the mean performance across the models from CV was used. The Figure 5.5, confirmed that the AE architecture surpassed the performance of the cGAN when using a small mono-region dataset.

For each model architecture, violin plots shown in Figure 5.5 and Table 5.7 illustrate the following:

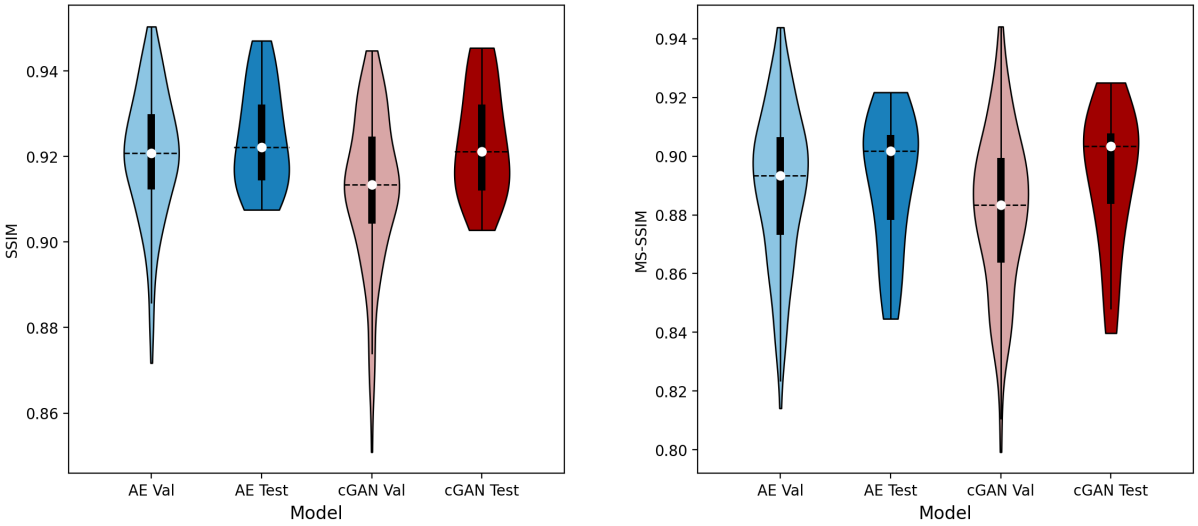
- The validation results confirmed that the AE exhibited the highest **PSNR** (28.36 ± 1.91 HU, with $p - value = 3.15 \times 10^{-8}$), while test results showed no statistically significant difference ($p - value = 0.06$).
- Both validation and test p-values (1.27×10^{-13} and below 0.01, respectively) supported the superior performance of the AE for the **MAE** in validation (69.54 ± 13.22 HU) and test (71.69 ± 14.28 HU).
- These plots show that on the validation set for the **SSIM** and **MS-SSIM**, the AE achieved

higher values (SSIM=0.92±0.01; MS-SSIM=0.89±0.03), supported by p-values 3.77×10^{-20} (SSIM) and 2.15×10^{-13} (MS-SSIM). However, these differences were not statistically significant in the test set results.



(a) PSNR: $p_{\text{val}} = 3.15 \times 10^{-8}$; $p_{\text{test}} = 0.06$

(b) MAE: $p_{\text{val}} = 1.27 \times 10^{-13}$; $p_{\text{test}} < 0.01$



(c) SSIM: $p_{\text{val}} = 3.77 \times 10^{-20}$; $p_{\text{test}} = 0.15$

(d) MS-SSIM: $p_{\text{val}} = 2.15 \times 10^{-13}$; $p_{\text{test}} = 0.25$

Figure 5.5: Violin plots comparing the image quality metrics between AE and cGAN best models. The statistically significant differences ($p < 0.05$) were investigated through a paired sample t-test. The bold line represents the interquartile range (IQR) (25%-75%), with the white dot marking the median (50%). The dashed horizontal line at 50% was added to facilitate visual comparison between the plots. Whiskers extend to $1.5 \times IQR$, and outliers are shown beyond this range.

Table 5.7: Test results of the ensemble of models using SynthRAD2023.

Model	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
AE	28.79 ± 1.81	71.69 ± 14.28	0.92 ± 0.01	0.89 ± 0.02
cGAN	28.99 ± 1.83	76.62 ± 13.40	0.92 ± 0.01	0.89 ± 0.02

5.2 SynthRAD2025

5.2.1 Centre-stratified approach

Both the AE and cGAN were trained and validated with 90% of data from SynthRAD2025, stratified by centres in order to ensure a proportional representation of each centre within each fold of the stratified 5-fold CV. For testing, the remaining 10% of SynthRAD2025 data was used to evaluate the models.

AE

As illustrated in Table 5.8, the average metrics achieved across all folds during CV were highly dependent on the anatomical region and medical centre. The results revealed for centre A that PSNR, SSIM and MS-SSIM metrics accomplished considerably better values for the HN region, whereas the best MAE (lower value) was achieved in AB region; centre B that images from TH region presented better metrics for PSNR, SSIM and MS-SSIM than the MAE metric, which was the lowest (better) when using images from the AB region; centre C that all metrics were superior for the AB region, except for the MS-SSIM that was better when using HN region.

Table 5.8: Stratified 5-fold cross validation results of image metrics for the generated sCT scans using the AE model trained in all centres of SynthRAD2025 dataset. The mean is calculated across each value from each case.

Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
A	AB	26.07 ± 0.53	102.49 ± 3.32	0.91 ± 0.01	0.88 ± 0.00
A	TH	25.78 ± 0.24	137.60 ± 4.33	0.92 ± 0.00	0.90 ± 0.00
A	HN	31.46 ± 0.22	117.55 ± 1.15	0.96 ± 0.00	0.95 ± 0.00
B	AB	28.27 ± 0.14	123.44 ± 2.92	0.93 ± 0.00	0.90 ± 0.00
B	TH	30.04 ± 0.05	125.25 ± 1.51	0.95 ± 0.00	0.94 ± 0.00
C	AB	29.16 ± 0.20	113.99 ± 3.52	0.94 ± 0.00	0.89 ± 0.00
C	HN	28.29 ± 0.06	170.15 ± 1.57	0.93 ± 0.00	0.92 ± 0.00
Mean	-	**28.45 ± 2.96**	**127.50 ± 33.18**	**0.93 ± 0.02**	**0.91 ± 0.04**

Concerning test results, the internally test set reported that: i) centre A PSNR, SSIM, and MS-SSIM' values where larger for the HN region, except for MAE that reached its best value for AB region; ii) centre B followed the same trend as the validation results, in which TH region input images generated better metrics for PSNR, SSIM, and MS-SSIM, and MAE was the lowest for AB region; iii) centre C showed greater results for the AB region in PSNR, MAE, and SSIM, where as the MS-SSIM was higher in HN region. When performing the evaluation on unseen data from an external centre, the results were even better than those from the internal test.

Table 5.9: Test results of an ensemble of AE models trained in all centres (centres ABC - 90%). The models were tested internally (centres ABC - 10%) with MRI scans across different anatomical regions and externally (centre D) with MRI scans from HN region, both from SynthRAD2025 data. The mean is calculated across each value from each case.

Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
A	AB	26.27 ± 3.35	102.82 ± 40.95	0.92 ± 0.03	0.90 ± 0.04
A	TH	27.59 ± 1.51	113.55 ± 18.75	0.94 ± 0.01	0.93 ± 0.02
A	HN	29.70 ± 3.14	123.85 ± 30.49	0.94 ± 0.03	0.94 ± 0.03
B	AB	28.39 ± 3.10	112.16 ± 23.44	0.93 ± 0.03	0.89 ± 0.04
B	TH	29.56 ± 1.57	122.56 ± 35.20	0.95 ± 0.01	0.93 ± 0.02
C	AB	29.03 ± 0.00	128.98 ± 0.00	0.95 ± 0.00	0.89 ± 0.00
C	HN	28.13 ± 1.24	186.56 ± 52.88	0.93 ± 0.00	0.91 ± 0.01
Mean	-	**28.42 ± 2.69**	**124.78 ± 41.13**	**0.94 ± 0.02**	**0.92 ± 0.03**
D	HN	**29.67 ± 1.49**	**111.69 ± 13.83**	**0.94 ± 0.01**	**0.94 ± 0.01**

cGAN

Both validation (Table 5.10) and test (Table 5.11) tables present the mean cGAN results across the 5 folds in a centre-based approach. In Table 5.10, centre A showed the best performing metrics for HN region, except for the MAE value that was better in AB region. Additionally, centre B MAE value was lower for the AB region, while the other metrics were superior for the TH region. The centre c results achieved higher performances in PSNR, MAE, and SSIM of the AB region, and a better mean value in MS-SSIM for the HN region. The results illustrated in Table 5.11 show that the model generalises well to new data, since a superior performance was achieved in the internal test compared to validation. The external test results were consistent with the internal ones, but showed slightly better performance.

Table 5.10: Stratified 5-fold cross validation results of image metrics for the generated sCT scans using the cGAN model trained in all centres of SynthRAD2025 dataset. The mean is calculated across each value from each case.

Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
A	AB	25.87 ± 0.57	104.53 ± 6.10	0.91 ± 0.01	0.88 ± 0.00
A	TH	25.50 ± 0.15	142.58 ± 4.66	0.92 ± 0.01	0.90 ± 0.00
A	HN	31.27 ± 015	121.25 ± 2.10	0.95 ± 0.00	0.95 ± 0.00
B	AB	28.97 ± 0.10	129.27 ± 1.89	0.92 ± 0.00	0.89 ± 0.00
B	TH	29.62 ± 0.08	132.76 ± 3.20	0.94 ± 0.00	0.93 ± 0.00
C	AB	28.83 ± 0.26	119.49 ± 2.93	0.94 ± 0.00	0.88 ± 0.00
C	HN	28.19 ± 0.06	173.45 ± 3.02	0.93 ± 0.00	0.92 ± 0.00
Mean	-	**28.07 ± 2.99**	**137.31 ± 35.33**	**0.92 ± 0.04**	**0.91 ± 0.04**

Table 5.11: Test results of an ensemble of cGAN models trained in all centres (centres ABC - 90%). The models were tested internally (centres ABC - 10%) with MRI scans across different anatomical regions and externally (centre D) with MRI scans from HN region, both from SynthRAD2025 data. The mean is calculated across each value from each case.

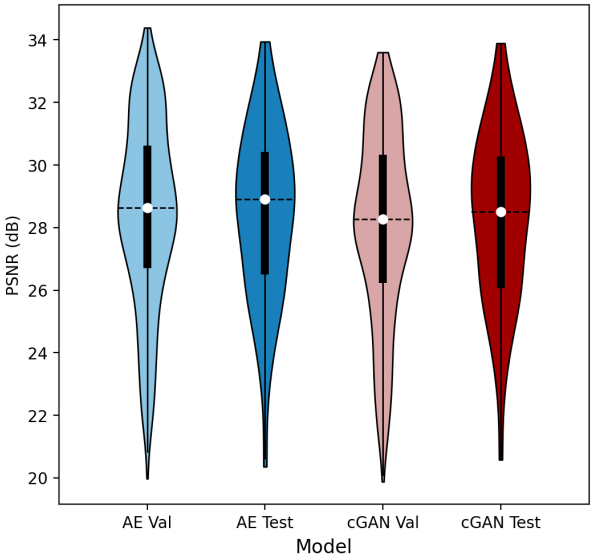
Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
A	AB	26.22 ± 3.22	104.33 ± 38.22	0.91 ± 0.03	0.90 ± 0.04
A	TH	27.35 ± 1.51	120.05 ± 21.83	0.93 ± 0.01	0.93 ± 0.02
A	HN	29.67 ± 3.10	126.28 ± 30.12	0.94 ± 0.03	0.94 ± 0.03
B	AB	28.24 ± 3.00	114.75 ± 22.16	0.92 ± 0.03	0.89 ± 0.04
B	TH	29.52 ± 1.61	124.92 ± 35.45	0.95 ± 0.01	0.93 ± 0.02
C	AB	29.44 ± 0.00	122.57 ± 0.00	0.95 ± 0.00	0.90 ± 0.00
C	HN	28.13 ± 1.19	188.32 ± 51.27	0.93 ± 0.00	0.92 ± 0.01
Mean	-	**28.34 ± 2.66**	**127.60 ± 40.46**	**0.93 ± 0.02**	**0.92 ± 0.03**
D	HN	**29.69 ± 1.52**	**116.09 ± 14.09**	**0.94 ± 0.01**	**0.94 ± 0.01**

5.2.2 Performance comparison: AE vs. cGAN

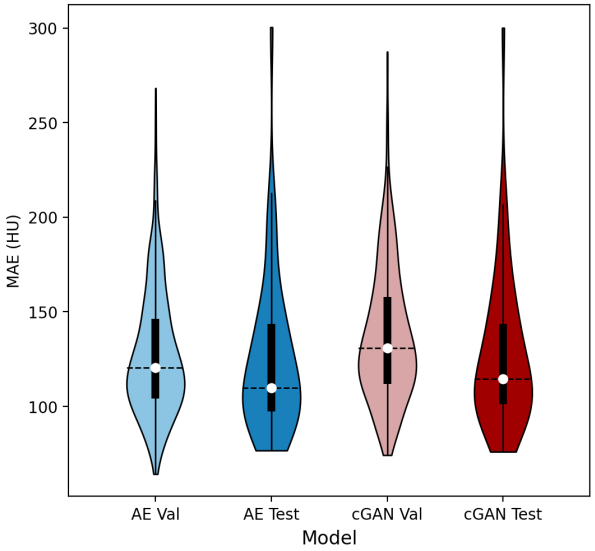
According to Figure 5.6, the result obtained in the previous dataset was reinforced: AE had superior performance in relation to the cGAN, this time utilising a larger, multi-region dataset.

In the validation set, the AE achieved a PSNR of 28.45 ± 2.96 HU, MAE of 127.50 ± 33.18 HU, and SSIM of 0.93 ± 0.02, with p-values: 4.54×10^{-32}, 6.32×10^{-37}, 7.18×10^{-14}, respectively. In the internal test set, the AE obtained a PSNR of 28.42 ± 2.69 dB, MAE of 124.78 ± 41.13 HU, and SSIM of 0.94 ± 0.02, with all p-values below 0.05 (p-values: < 0.01, 1.25×10^{-5}, 1.26 ± 10^{-7}, respectively).

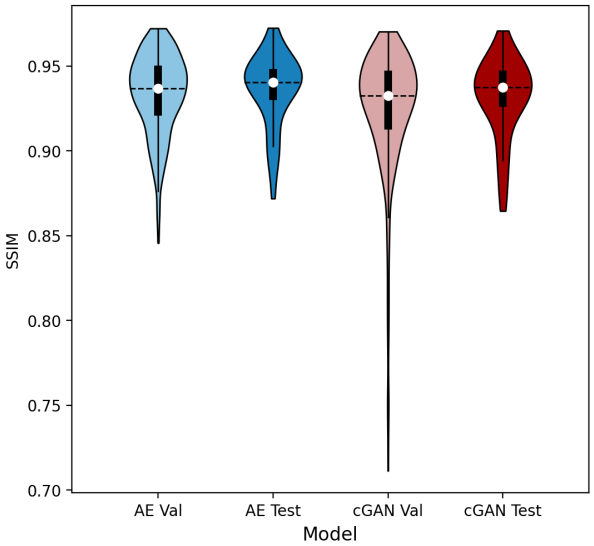
For MS-SSIM, validation results showed a statistically significant difference of 0.91 ± 0.04(p-value=5.31×10^{-18}), in which AE reached hardly better values than cGAN, only noticeable due to the dashed line in the violin plots. Nevertheless, differences in MS-SSIM test results were not statistically significant ($p - value = 0.68$).



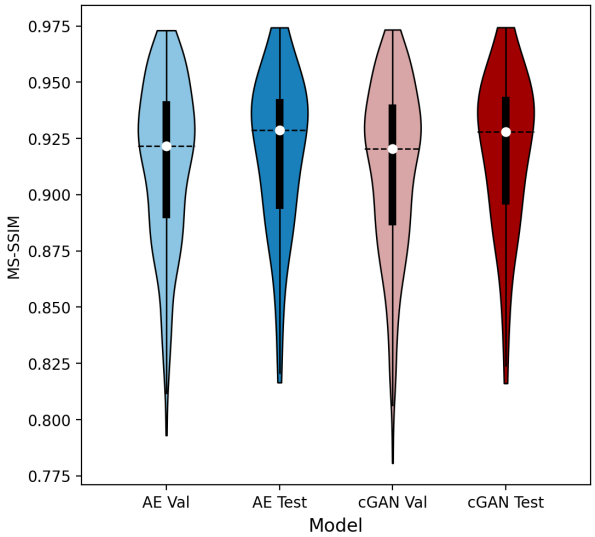
(a) PSNR: $p_{\text{val}} = 4.54 \times 10^{-32}$; $p_{\text{test}} < 0.01$



(b) MAE: $p_{\text{val}} = 6.22 \times 10^{-37}$; $p_{\text{test}} = 1.25 \times 10^{-5}$



(c) SSIM: $p_{\text{val}} = 7.18 \times 10^{-14}$; $p_{\text{test}} = 1.26 \times 10^{-7}$



(d) MS-SSIM: $p_{\text{val}} = 5.31 \times 10^{-18}$; $p_{\text{test}} = 0.68$

Figure 5.6: Violin plots comparing the image quality metrics obtained during validation and testing (except for data from centre D) with the ensemble of models for AE and cGAN best models. The statistically significant differences ($p < 0.05$) were investigated using the paired sample t-test. The bold line represents the interquartile range (IQR) (25%-75%), with the white dot marking the median (50%). The dashed horizontal line at the median was added to facilitate visual comparison between the plots. Whiskers extend to $1.5 \times IQR$, and outliers are shown beyond this range.

5.2.3 Region-based approach

The performance of the AE generative model was evaluated for models trained only in individual anatomical regions (AB, TH, HN) and all regions combined. Then, a comparison was

made between the two approaches (one-region and multi-region) using test data from the same region to identify which one generated the highest-quality sCTs.

Models trained on specific single regions

According to Table 5.12, the results show that the model trained using 5-fold cross validation exclusively on HN region data achieved the highest mean PSNR, SSIM, and MS-SSIM values, while the best (lower) MAE mean value was obtained by the model trained exclusively on AB region data. In particular, the results from centre C in the model trained only on HN images were significantly higher than those obtained with the model trained only on AB region for the same centre. Although the model trained on AB region alone achieved the best MAE mean metric (117.23 ± 33.54 HU), it yielded the worst mean values for the the remaining metrics among all models.

Table 5.12: 5-fold cross validation results of image metrics for the generated sCT scans using the AE models, each trained exclusively on one region data: AB, TH, HN, from the SynthRAD2025 dataset. Metrics are reported per centre and aggregated (across each value from each case) for each model.

Region	Centre	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
Model trained on AB region group only					
AB	A	25.80 ± 0.40	105.33 ± 3.72	0.91 ± 0.00	0.88 ± 0.00
AB	B	27.57 ± 0.38	126.98 ± 2.15	0.92 ± 0.00	0.89 ± 0.00
AB	C	28.25 ± 0.61	116.20 ± 4.23	0.94 ± 0.00	0.88 ± 0.01
Mean	-	**27.23 ± 2.87**	**117.23 ± 33.54**	**0.92 ± 0.02**	**0.89 ± 0.03**
Model trained on TH region group only					
TH	A	26.25 ± 0.09	129.57 ± 2.79	0.92 ± 0.00	0.91 ± 0.00
TH	B	29.70 ± 0.16	127.77 ± 1.41	0.95 ± 0.00	0.94 ± 0.00
Mean	-	**27.95 ± 3.02**	**131.09 ± 34.34**	**0.93 ± 0.02**	**0.92 ± 0.03**
Model trained on HN region group only					
HN	A	31.68 ± 0.19	114.09 ± 3.67	0.95 ± 0.00	0.96 ± 0.00
HN	C	28.16 ± 0.11	173.85 ± 1.64	0.93 ± 0.00	0.91 ± 0.00
Mean	-	**30.55 ± 2.09**	**134.04 ± 30.90**	**0.95 ± 0.02**	**0.94 ± 0.03**

Test results per model shown in Table 5.13, revealed that the model trained solely in HN data achieved the best mean PSNR value (29.10 ± 2.66 dB), while the model trained in AB data alone achieved the best MAE value (112.54 ± 29.58 HU). The highest mean values for SSIM and MS-SSIM were obtained for both models trained exclusively in TH and HN regions. Similarly to the validation results, centre C from the HN region data had the worst MAE value (186.69 ± 49.82 HU), which influenced the model's internal test results. However, the external test results from centre D were superior to those from the internal test.

Table 5.13: Test results of image metrics for the generated sCT scans using the ensemble of AE models, each trained exclusively on one region data: AB, TH, HN, from the SynthRAD2025 dataset. Metrics are reported per centre and aggregated (across each value from each case) for each model. For the HN region, results include both internal test (10% hold-out test set with the same centres used for training), and external test, on unseen centre D data.

Region	Centre	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
Model trained on AB region group only					
AB	A	26.08 ± 2.92	106.41 ± 38.30	0.92 ± 0.03	0.90 ± 0.04
AB	B	28.24 ± 2.95	115.43 ± 38.30	0.93 ± 0.03	0.89 ± 0.04
AB	C	29.38 ± 0.00	123.37 ± 0.00	0.96 ± 0.00	0.90 ± 0.00
Mean	-	**27.50 ± 3.06**	**112.54 ± 29.58**	**0.93 ± 0.03**	**0.89 ± 0.04**
Model trained on TH region group only					
TH	A	27.56 ± 1.56	112.50 ± 19.21	0.94 ± 0.01	0.93 ± 0.02
TH	B	29.13 ± 1.50	129.07 ± 35.39	0.95 ± 0.01	0.93 ± 0.02
Mean	-	**28.34 ± 1.72**	**120.78 ± 29.66**	**0.94 ± 0.01**	**0.93 ± 0.02**
Model trained on HN region group only					
HN	A	29.72 ± 3.15	125.07 ± 29.91	0.94 ± 0.03	0.94 ± 0.04
HN	C	28.17 ± 1.20	186.69 ± 49.82	0.93 ± 0.00	0.92 ± 0.01
Mean	-	**29.10 ± 2.66**	**149.71 ± 49.41**	**0.94 ± 0.02**	**0.93 ± 0.03**
HN	D	29.68 ± 1.54	116.19 ± 14.19	0.94 ± 0.01	0.94 ± 0.01

Model trained on all regions

Table 5.14 presents the validation results in 5 folds, indicating that throughout all metrics,

images from TH region yielded the best values for centre B, while the HN region performance excelled in centre A. Regarding the results for the AB region, the best PSNR and SSIM values were observed in centre C, the lowest MAE in centre A, and the highest MS-SSIM in centre B.

Table 5.14: 5-fold cross validation results of image metrics for the generated sCT scans using the AE model trained in all regions of SynthRAD2025 dataset. Mean is calculated across each value from each case.

Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
A	**AB**	26.14 ± 0.33	100.54 ± 4.27	0.92 ± 0.01	0.88 ± 0.00
B	**AB**	28.43 ± 0.15	116.24 ± 2.04	0.93 ± 0.00	0.90 ± 0.00
C	**AB**	29.65 ± 0.88	110.42 ± 6.01	0.95 ± 0.00	0.89 ± 0.01
A	**TH**	26.18 ± 0.12	131.48 ± 2.64	0.93 ± 0.00	0.91 ± 0.00
B	**TH**	30.07 ± 0.15	123.92 ± 1.18	0.95 ± 0.00	0.94 ± 0.00
A	**HN**	31.73 ± 0.33	115.47 ± 1.14	0.96 ± 0.00	0.96 ± 0.00
C	**HN**	28.87 ± 0.10	156.99 ± 1.59	0.93 ± 0.00	0.92 ± 0.00
-	**Mean**	**28.41 ± 2.91**	**116.93 ± 30.23**	**0.93 ± 0.02**	**0.91 ± 0.04**

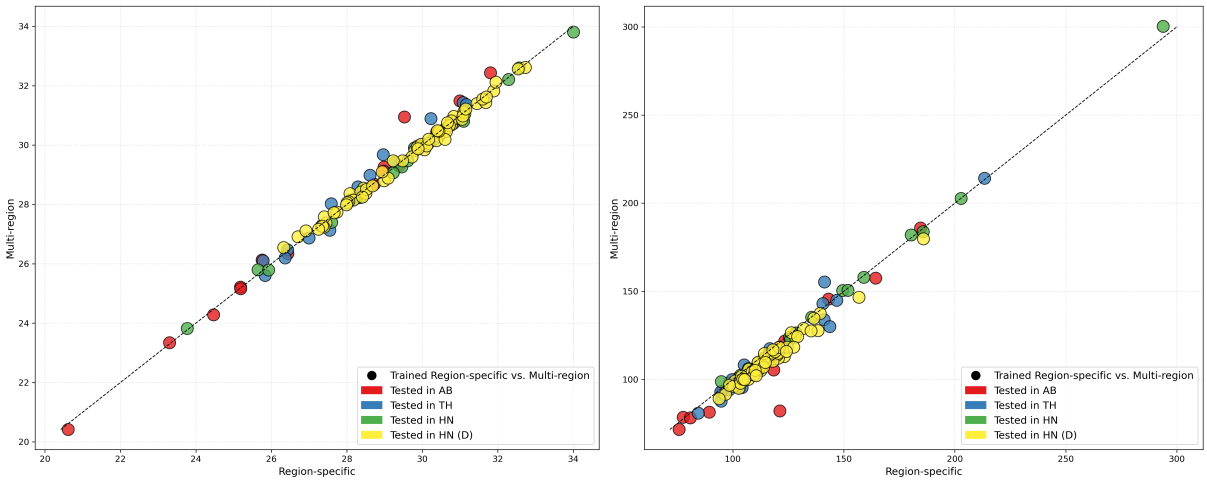
Results from Table 5.15 summarise the performance in five different test sets using the multi-region model (model trained on data from all regions). Among testes on different sets of images, the highest mean PSNR value (29.01 ± 2.62 dB) was observed for the HN region, while the best mean MAE value (106.99 ± 30.93 HU) was in the AB region. The external test (centre D) achieved better mean metrics than those from the internal test (centres A, B, C) with HN data, except for the MAE metric (111.52 ± 13.83 HU) and the SSIM metric (0.94 ± 0.01).

Table 5.15: Test results of image metrics for the generated sCT scans using the ensemble of AE models trained on all regions (regions AB, TH and HN - 90%) and tested on the remaining data from all regions (regions AB, TH and HN - 10%), as well as on individual regions, from the SynthRAD2025 dataset. Metrics are reported per centre and aggregated (across each value from each case) for each test. For the HN region, results include both internal test (10% hold-out test set with the same centres used for training), and external test, on unseen centre D data.

Centre	Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
Tested on AB+TH+HN data					
A	**AB**	26.35 ± 3.30	99.08 ± 39.65	0.92 ± 0.03	0.91 ± 0.04
B	**AB**	28.39 ± 3.11	110.64 ± 23.94	0.93 ± 0.03	0.89 ± 0.04
C	**AB**	29.30 ± 0.00	121.77 ± 0.00	0.95 ± 0.00	0.90 ± 0.00
A	**TH**	27.50 ± 1.58	113.73 ± 22.86	0.94 ± 0.01	0.93 ± 0.02
B	**TH**	29.48 ± 1.56	123.19 ± 37.25	0.95 ± 0.01	0.93 ± 0.02
A	**HN**	29.64 ± 3.08	123.95 ± 29.92	0.94 ± 0.03	0.94 ± 0.03
C	**HN**	28.07 ± 1.21	187.54 ± 52.25	0.93 ± 0.00	0.91 ± 0.01
-	**Mean**	**28.39 ± 2.67**	**124.19 ± 42.07**	**0.94 ± 0.02**	**0.92 ± 0.03**
Tested only on AB data					
A	**AB**	26.35 ± 3.30	99.08 ± 39.65	0.92 ± 0.03	0.91 ± 0.04
B	**AB**	28.39 ± 3.11	110.64 ± 23.94	0.93 ± 0.03	0.89 ± 0.04
C	**AB**	29.30 ± 0.00	121.77 ± 0.00	0.95 ± 0.00	0.90 ± 0.00
-	**Mean**	**27.68 ± 3.26**	**106.99 ± 30.93**	**0.93 ± 0.03**	**0.90 ± 0.04**
Tested only on TH data					
A	**TH**	27.50 ± 1.58	113.73 ± 22.86	0.94 ± 0.01	0.93 ± 0.02
B	**TH**	29.48 ± 1.56	123.19 ± 37.25	0.95 ± 0.01	0.93 ± 0.02
-	**Mean**	**28.49 ± 1.85**	**118.46 ± 31.27**	**0.94 ± 0.01**	**0.93 ± 0.02**
Tested only on HN data					
A	**HN**	29.64 ± 3.08	123.95 ± 29.92	0.94 ± 0.03	0.94 ± 0.03
C	**HN**	28.07 ± 1.21	187.54 ± 52.25	0.93 ± 0.00	0.91 ± 0.01
-	**Mean**	**29.01 ± 2.62**	**149.39 ± 50.99**	**0.94 ± 0.02**	**0.93 ± 0.03**
Tested only on HN external data					
D	**HN**	29.66 ± 1.50	111.52 ± 13.83	0.94 ± 0.01	0.94 ± 0.01

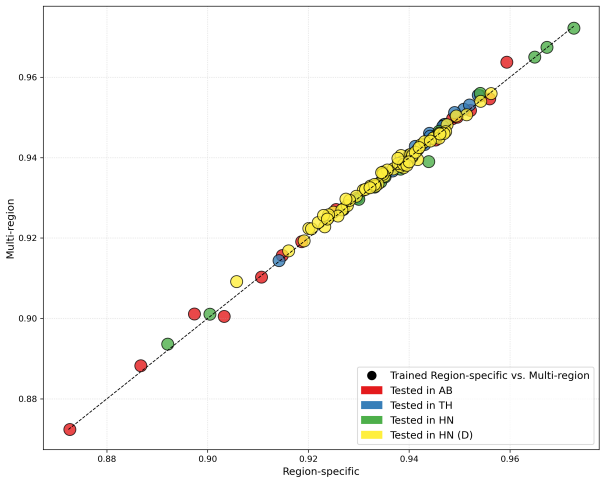
One-region model vs. Multi-region model

From Figure 5.7, the analysis is based on the distribution of the points relative to the $y = x$ line. If the majority of the points are above the line ($y > x$), this indicates that the multi-region model outperformed the region-specific models. This trend is applied to PSNR, SSIM, and MS-SSIM graphs. Conversely, for the MAE, which is an error that we aim to minimise, if the majority of the circles are below the line ($y < x$), the multi-region outperformed the individual region models. Taking this into account, Figure 5.7a shows a relatively balanced distribution of points along the line, whereas the difference in the other metrics is more pronounced. In the Figure 5.7c, Figure 5.7d, the multi-region model achieves better test results, since there are more circles above the diagonal. Similarly, in Figure 5.7b, the majority of circles lie below the line, which also indicates superior performance of the multi-region model.

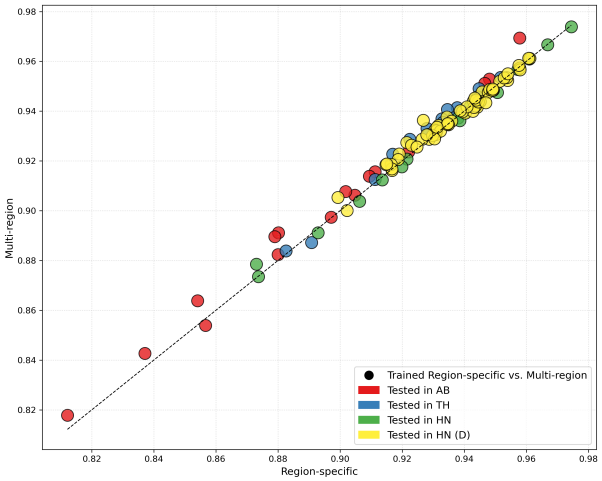


(a) PSNR (dB)

(b) MAE (HU)



(c) SSIM



(d) MS-SSIM

Figure 5.7: Comparison between the test results of models trained in a specific region versus a model trained in all regions. Each model trained in AB, TH, or HN images was tested on data from the same region. The model trained in data from all regions was tested in each individual region.

The subsequent tables present the p-values of the paired-sample t-tests between metric values for a single-region model (model trained on individual region) and a multi-region model (model trained on all regions). The metric values are the mean of the test results across all centres (from: Table 5.15).

The metrics reported in Table 5.16, show statistically significant differences for MAE ($p_{\text{value}} = 0.0385$) and MS-SSIM ($p_{\text{value}} = 1.5602 \times 10^{-4}$), in which the multi-region model had the best values, even though the difference observed in MS-SSIM was minimal (0.8936 vs. 0.8937). Results from Table 5.17 indicate that all metrics, except for the PSNR, showed statistically significant differences, which consistently indicated better performance for the multi-region model.

Table 5.16: Predictions on AB test set: comparison between models trained with AB images and all regions images.

Metric	p-value	AB model	AB+TH+HN model
PSNR (dB,↑)	0.1012	27.5029	27.6814
MAE (HU,↓)	0.0385*	112.5432	**106.9976**
SSIM (↑)	0.1820	0.9271	0.9278
MS-SSIM (↑)	1.5602×10^{-4}*	0.8936	**0.8937**

Table 5.17: Predictions on TH test set: comparison between models trained with TH images and all regions images.

Metric	p-value	TH model	AB+TH+HN model
PSNR (dB,↑)	0.0535	28.3417	28.4899
MAE (HU,↓)	0.1134*	120.7821	**118.4613**
SSIM (↑)	1.1149×10^{-5}*	0.9425	**0.9435**
MS-SSIM (↑)	8.3197×10^{-4}*	0.9284	**0.9309**

According to Table 5.18, only PSNR mean differences were statistically different ($p_{\text{value}} = 0.096$), with HN model achieving the best results. Apart from PSNR ($p_{\text{value}} = 0.2607$), the external testing of individual HN and multi-region models (Table 5.18) revealed statistically significant differences that rejected the null hypothesis, and indicated a better performance of the multi-region models.

Table 5.18: Predictions on HN test set and HN external centre D test set: comparison between models trained with HN images and all regions images.

Metric	p-value	HN model	AB+TH+HN model
Internal test			
PSNR (dB,↑)	0.0096*	**29.1008**	29.0103
MAE (HU,↓)	0.6645	149.7142	149.3876
SSIM (↑)	0.7339	0.9385	0.9384
MS-SSIM (↑)	0.1809	0.9286	0.9279
External test			
PSNR (dB,↑)	0.2607	29.6790	29.6605
MAE (HU,↓)	5.6009×10^{-23}*	116.1901	**111.5245**
SSIM (↑)	4.9472×10^{-4}*	0.9355	**0.9360**
MS-SSIM (↑)	0.0158*	0.9373	**0.9379**

Having established that the multi-region model generally performs better than the one-region model, the images below illustrate the worst (Figure 5.8) and best (Figure 5.9) cases in each region. These cases were evaluated using 3D masked volumes in HU, limited to the region of interest and normalised to a fixed data range of 4000 HU. Moreover, the images on the right show the pixel-wise difference between the synthetic-CT and the ground truth CT, in which the colour white indicates no difference. From the HN images, bone structures appear to be poorly generated, even in the best cases.

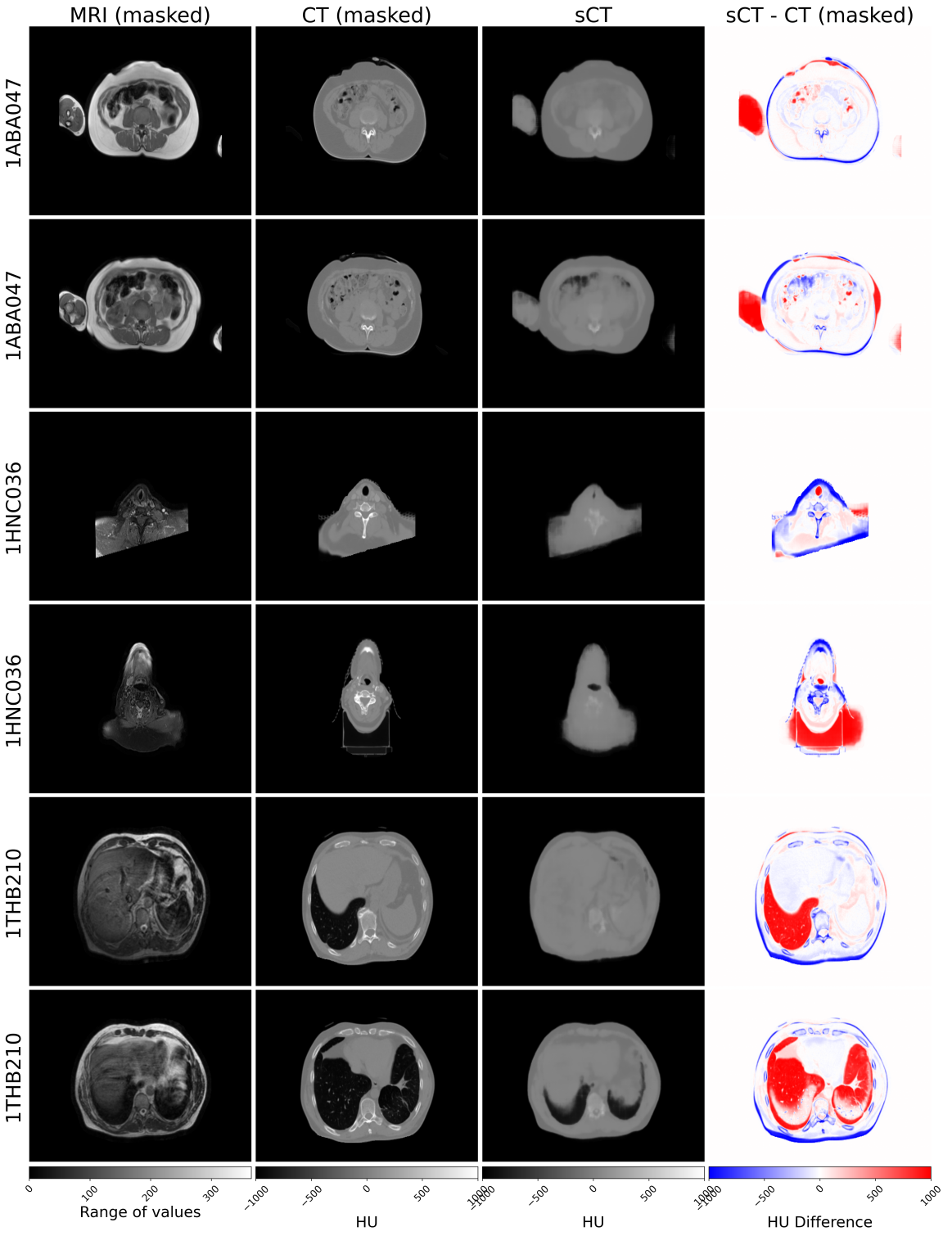


Figure 5.8: Slices examples of the worst mean ranked patients per region. *1ABA047*: PSNR=21.19 dB, MAE=185.90 HU, SSIM=0.64, MS-SSIM=0.55; *1HNC036*: PSNR=18.78 dB, MAE=300.36 HU, SSIM=0.44, MS-SSIM=0.54; *1THB210*: PSNR=20.70 dB, MAE=214.11 HU, SSIM=0.54, MS-SSIM=0.60.

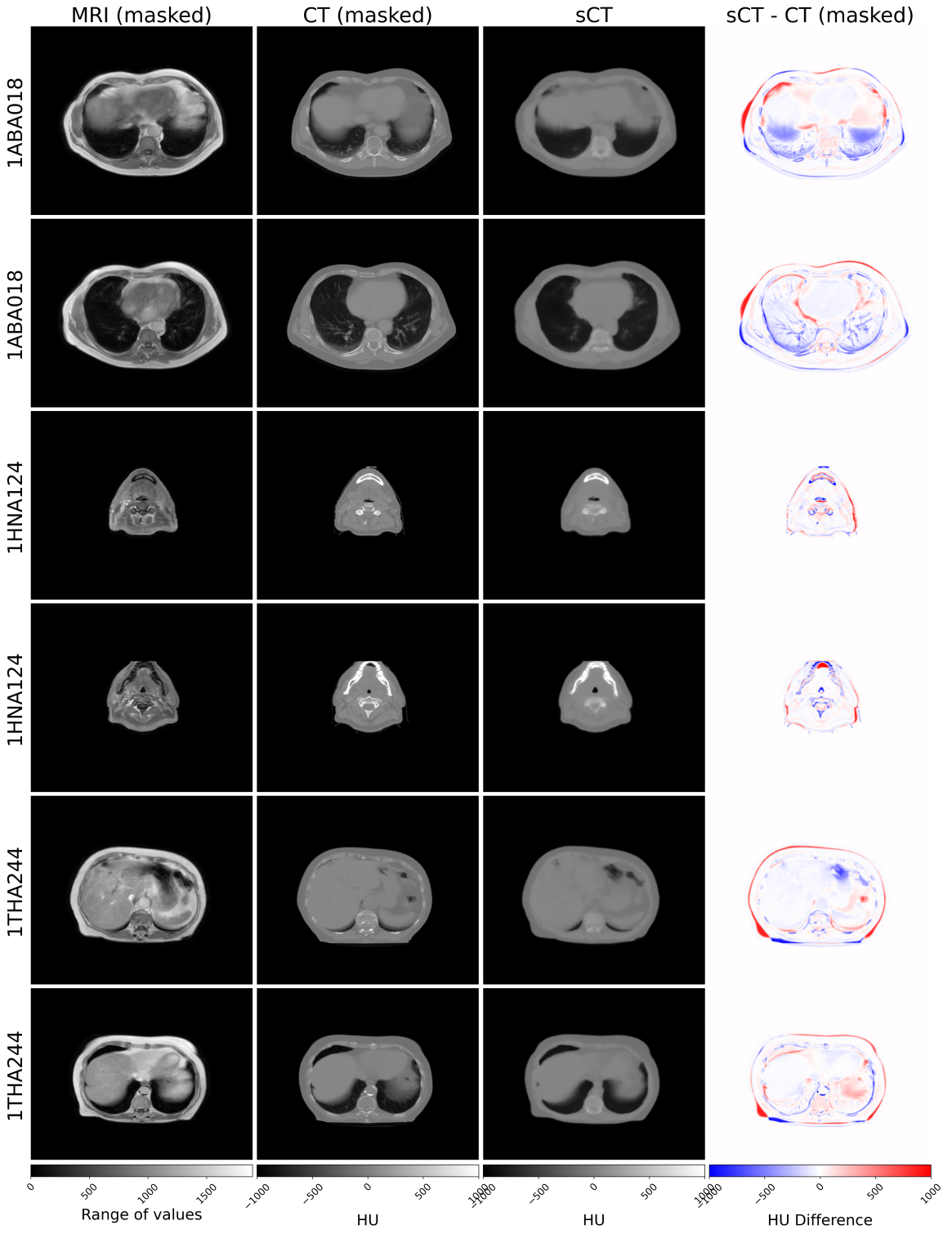


Figure 5.9: Slices examples of the best mean ranked patients per region. *1ABA018*: PSNR=28.60 dB, MAE=78.59 HU, SSIM=0.83, MS-SSIM=0.91; *1HNA124*: PSNR=26.11 dB, MAE=98.86 HU, SSIM=0.80, MS-SSIM=0.89; *1THA244*: PSNR=27.48 dB, MAE=80.94 HU, SSIM=0.81, MS-SSIM=0.90

5.2.4 Results for literature comparison

The following table, Table 5.19, includes the mean image quality metrics across all patients calculated for the **AE best ensemble model** using SynthRAD2023 test set and the **AE multi-region ensemble model** using SynthRAD2025 test set. The data presented are further analysed and compared with the existing literature in the Discussion chapter, as the metrics were calculated under comparable conditions to those in the literature. No statistically significant difference was found in the metrics between the centres of pelvis data. The test results obtained from SynthRAD2025 show superior performance for the AB region set in terms of PSNR, MAE, and SSIM. The highest MS-SSIM, even though with a marginal difference from the other region sets, was observed for the TH set.

Table 5.19: Mean test results across all patients per region for masked 3D volumes in HU. Metrics were calculated on 3D images in HU, limited to the region of interest and using a consistent data range of 4000 HU.

Region	PSNR (dB,↑)	MAE (HU,↓)	SSIM (↑)	MS-SSIM (↑)
SynthRAD2023 test set				
Pelvis	28.09 ± 1.51	71.69 ± 14.28	0.84 ± 0.03	0.86 ± 0.03
SynthRAD2025 test set				
AB	25.69 ± 1.99	106.99 ± 30.93	0.75 ± 0.06	0.79 ± 0.09
TH	25.06 ± 1.66	118.46 ± 31.27	0.72 ± 0.06	0.81 ± 0.07
HN	23.58 ± 2.04	149.40 ± 50.99	0.67 ± 0.09	0.80 ± 0.09

5.2.5 Geometric consistency metrics

Table 5.20 shows the list of all anatomical structures per region, which were divided into: i) the successfully segmented structures in both sCT and CT, which were used to calculate the mean metrics per region (Figure 5.10 and Figure 5.11) and the mean metrics per segment per region (Figure 5.12, Figure 5.13, and Figure 5.14); ii) the remaining expected structures that were not segmented. The tool failed to segment several bone structures from the sCT scans. Patients from HN region, had their vertebrae T1 to T6 included in CT scans when acquired at centre A, while scans at centre C had a limited field of view, which excluded all thoracic vertebrae.

As shown in Figure 5.10, the soft tissue segments in the AB region exhibited the highest mean mDICE values, while the bone segments were best generated in the HN region. The lowest value of mean HD95 (Figure 5.11), indicating superior performance, was obtained for soft tissues in the AB region and for bones in the HN region. However, the poorest segmentation performance, was observed in the TH and AB bone segments with the lowest mean mDICE value and in the HN soft tissue segments. Similarly, the highest mean HD95, which indicates poorer performance, was achieved in both TH bone and soft tissues segments.

Table 5.20: Per region, the structures segmented and those expected but not segmented by the TotalSegmentator tool in both sCT and CT. A structure is considered segmented only if it is identified in both images.

Region	Segmented structures	Not segmented structures
AB	Left and right kidneys, liver, stomach, all lung lobes, vertebrae S1, L1-L5, T1-T12, C7, heart, spinal cord, all ribs, and sternum	Vertebrae C1, C2, C3, C4, C5, C6
TH	Left and right kidneys, liver, stomach, all lung lobes, vertebrae L1-L5, T1-T12, C5-C7, heart, spinal cord, all right ribs, 1-11 left ribs and sternum	Vertebrae C1, C2, C3, C4, S1 and left 12th rib
HN	Oesophagus, trachea, thyroid, vertebrae C1-C7, and T1–T6, spinal cord, brain, skull	Vertebrae T7-T12

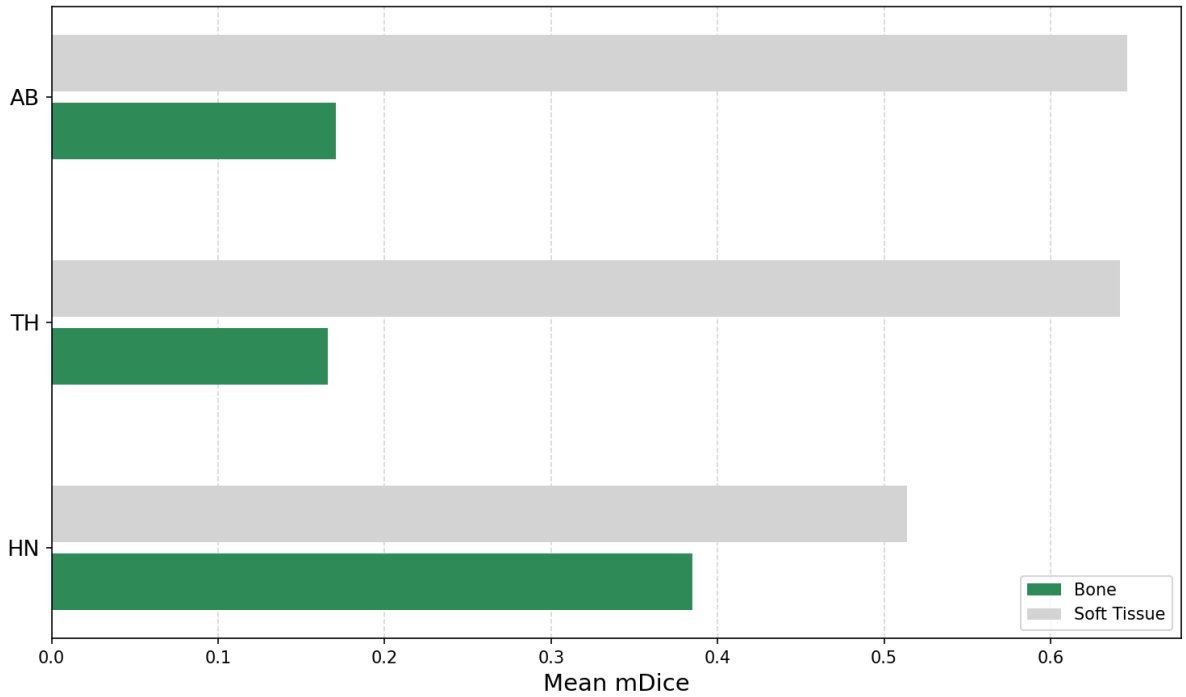


Figure 5.10: mDICE results averaged across all bone and soft tissue segments and patients within each region. Mean mDICE: 0.65 (AB - soft tissues), 0.17 (AB - bones), 0.64 (TH - soft tissues), 0.17 (TH - bones), and 0.51 (HN - soft tissues), 0.38 (HN - bones).

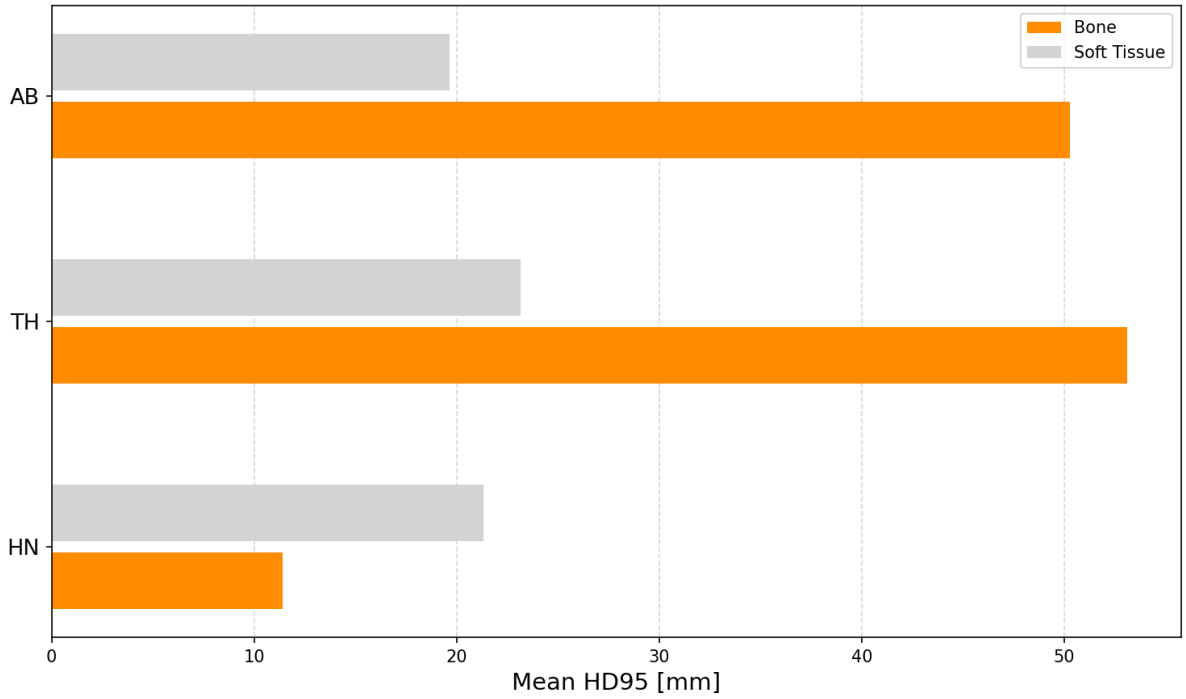


Figure 5.11: HD95 results averaged across all bone and soft tissue segments and patients within each region. Mean HD95 (mm): 19.65 (AB - soft tissues), 50.28 (AB - bones), 23.17 (TH - soft tissues), 53.10 (TH - bones), and 21.32 (HN - soft tissues), 11.41 (HN - bones).

The plots below illustrate that sCT segments resemble CT ones, when mDICE is close to 1 and HD95 is near 0, as observed, for example, in the liver in AB region, upper left lobe in TH region, and brain in HN region. It is important to note that not all patients got the same structures segmented, even if they were from the same region.

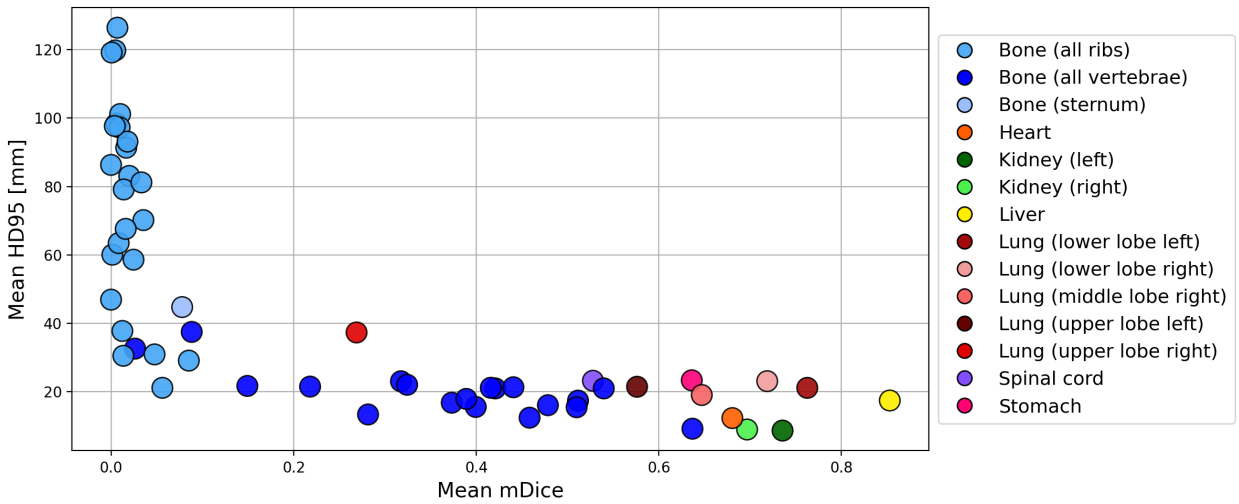


Figure 5.12: Mean metrics for each segment across all patients from AB region.

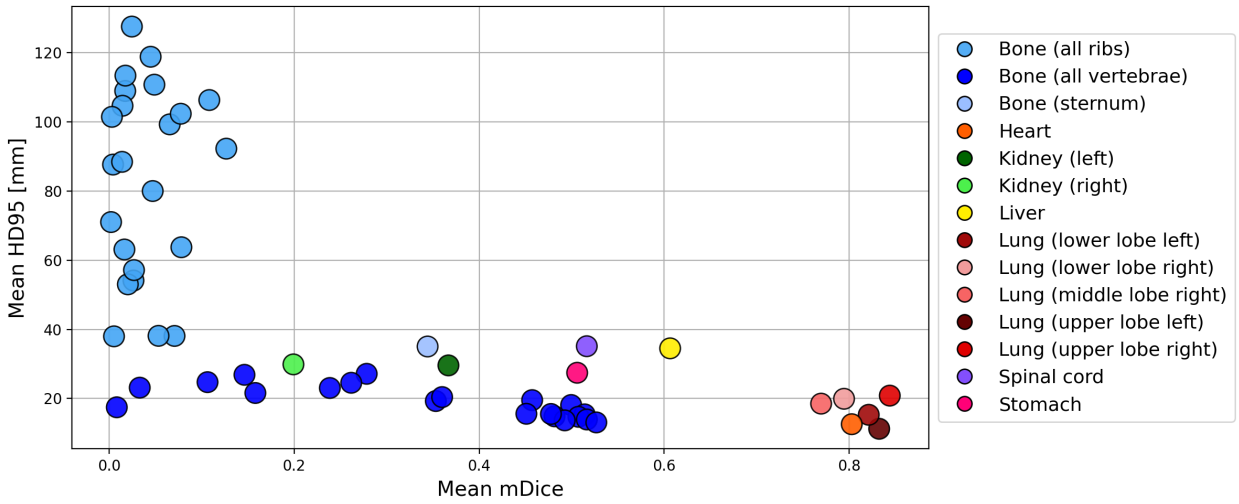


Figure 5.13: Mean metrics for each segment across all patients from TH region.

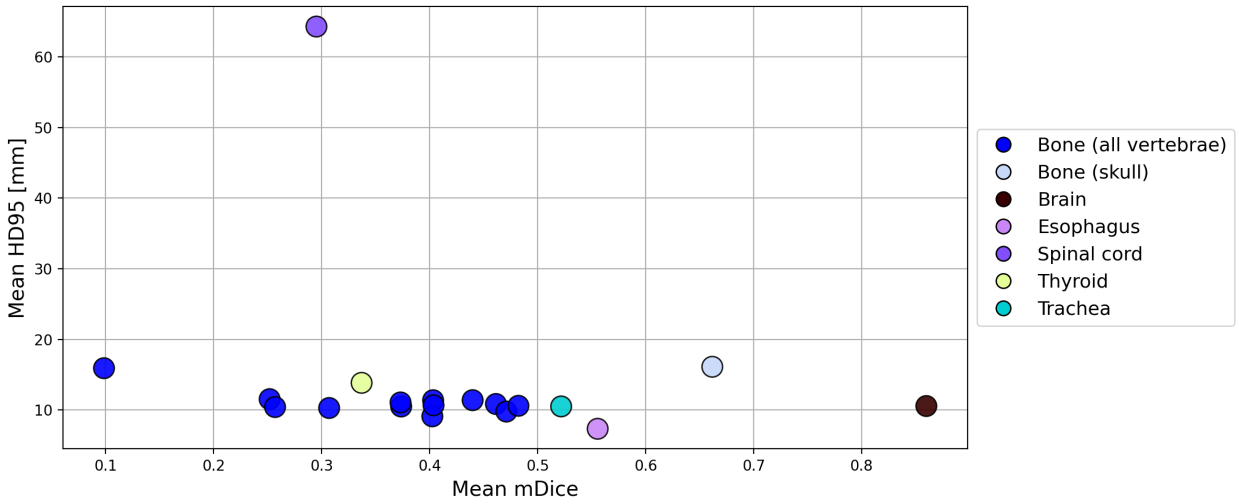


Figure 5.14: Mean metrics for each segment across all patients from HN region.

5.2.6 Dose metrics

The following table 5.21 shows a lower mean MAE value and higher mean gamma pass rates for the HN region. The lowest mean DVH metric was achieved in both AB and HN regions.

The division of the centres was performed as follows:

- AB region (16 patients): 6 patients (centre A), 9 patients (centre B), and 1 patient (centre C);
- TH region (18 patients): 9 patients (centre A), and 9 patients (centre B);

- HN region (15 patients): 9 patients (centre A), and 6 patients (centre C).

As such, the results by centres demonstrated for centre A, that the DVH and gamma pass rates values were consistently better across all different anatomical regions. The exceptional metric was the $MAE_{\text{target dose}}$, where centre B achieved superior results in the AB region. In TH region, however, centre A clearly yielded the best results. Although both centres for the HN region reported the same result (0.01 due to rounding) for this metric, the actual values were 0.007 for centre A and 0.008 for centre B. This indicates that centre A had superior results, even though the difference is minimal.

The treatment plans of patient 1THA028 were optimised using the same treatment parameters as the others. However, from the beam's "eye view", there was an overlap between the GTV (right upper lobe) and an OAR, which means that the optimisation on the sCT involved irradiating the OAR at risk. For that reason, this case represents an outlier compared to other results and was excluded from the TH analysis. The outlier's DVH is represented in Figure 5.18 and its dose distribution in Figure 5.20.

Table 5.21: Mean dose metrics acquired per region and centre from the proton treatment plan. 95% confidence intervals for mean gamma pass rates per region: [88.72, 96.27] (AB), [62.61, 76.14] (TH), [97.94, 99.23] (HN).

Region	Centre	$MAE_{\text{target dose}}$ (Gy, ↓)	DVH (↓)	$\gamma_{2\%/2\text{mm}}$ (%, ↑)
AB	A	0.03±0.03	0.27±0.15	93.79±3.69
AB	B	0.02±0.01	0.38±0.34	91.89±8.52
AB	C	0.06±0.00	0.60±0.00	90.20±0.00
Mean	-	**0.03±0.02**	**0.35±0.29**	**92.50±6.86**
TH	A	0.06±0.05	0.56±0.32	70.94±14.72
TH	B	0.14±0.22	0.67±0.46	67.98±10.53
Mean	-	**0.10±0.17**	**0.62±0.40**	**69.37±12.76**
HN	A	0.01±0.00	0.30±0.09	98.74±0.94
HN	C	0.01±0.00	0.44±0.42	98.35±1.32
Mean	-	**0.01±0.00**	**0.35±0.29**	**98.58±1.12**

Figures 5.15, 5.16, and 5.17 show the DVHs for the cases that achieved the lowest DVH

metric per region. According to these, the DVH curves for treatment plans based on CT and sCT almost completely overlap across all structures, indicating that treatment planning is independent of the type of image used. Figure 5.19 illustrates the best gamma pass rate case among all patients and regions, where the GTV is composed of the tongue, hard and soft palate, which are irradiated with higher doses (in red).

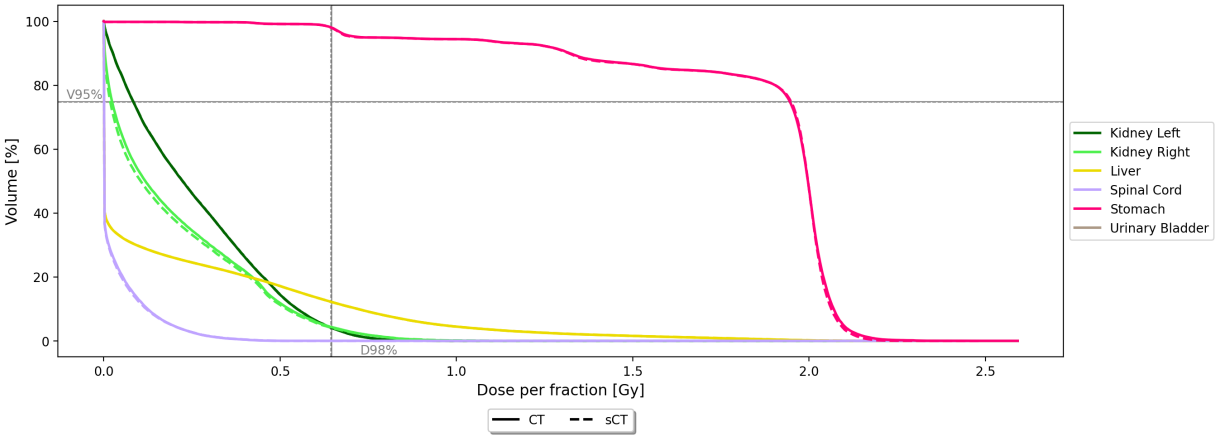


Figure 5.15: AB region, patient 1ABB059 (DVH_{metric}=0.02). The GTV is the stomach and the remaining structures are OARs. The gray lines indicate the dose that at least 98% of the target volume received (vertical axis), and the target volume that received at least 95% of the prescribed dose (horizontal axis).

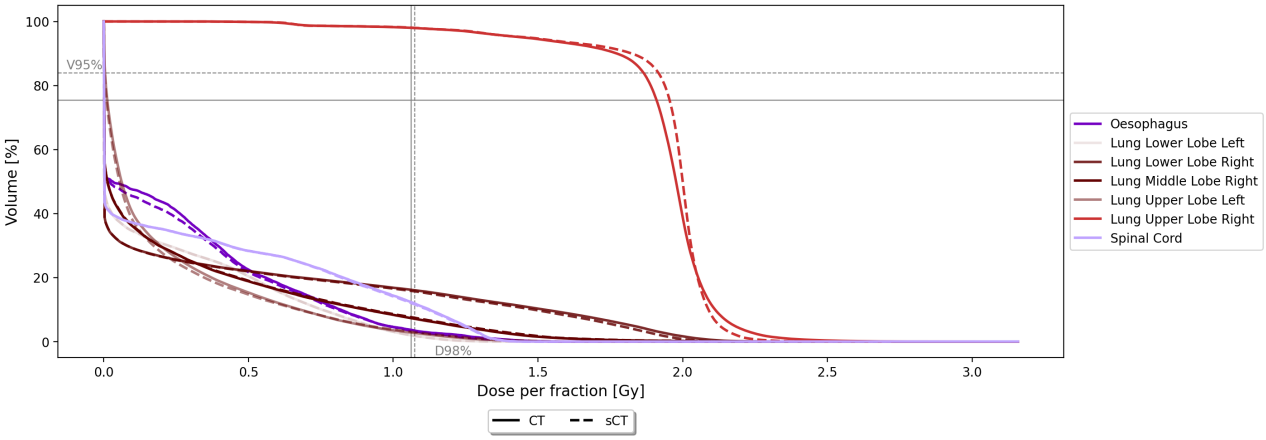


Figure 5.16: TH region, patient 1THB202 (DVH_{metric}=0.15). The GTV is the upper right lobe of the lung and the remaining structures are OARs. The gray lines indicate the dose that at least 98% of the target volume received (vertical axis), and the target volume that received at least 95% of the prescribed dose (horizontal axis).

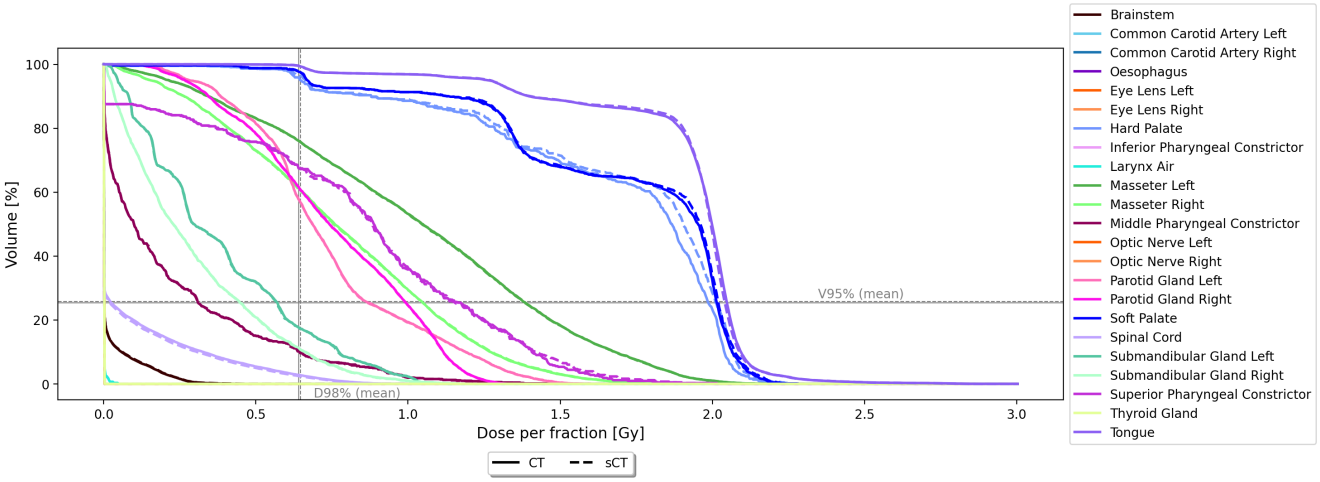


Figure 5.17: HN region, patient 1HNA085 ($DVH_{\mathbf{metric}}$=0.13). The GTV includes three structures: the tongue, hard and soft palates. The remaining structures are OARs. The gray lines indicate the dose that at least 98% of the target volume received (vertical axis), and the target volume that received at least 95% of the prescribed dose (horizontal axis). For the HN region, the gray lines are the values averaged across the 3 target structures that compose the GTV.

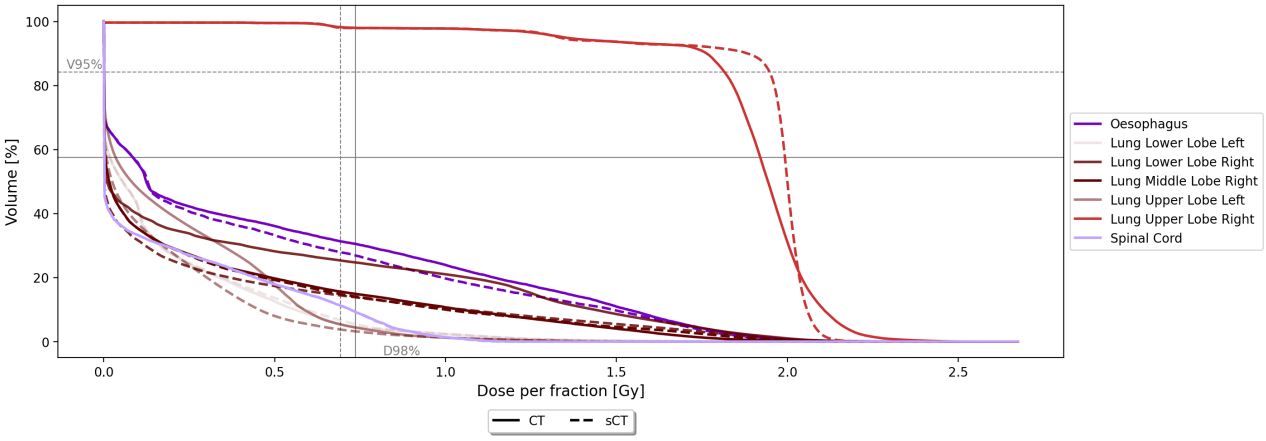


Figure 5.18: TH region, outlier patient 1THA028 ($DVH_{\mathbf{metric}} = 0.64$). The GTV is the upper right lobe of the lung and the remaining structures are OARs. The gray lines indicate the dose that at least 98% of the target volume received (vertical axis), and the target volume that received at least 95% of the prescribed dose (horizontal axis).

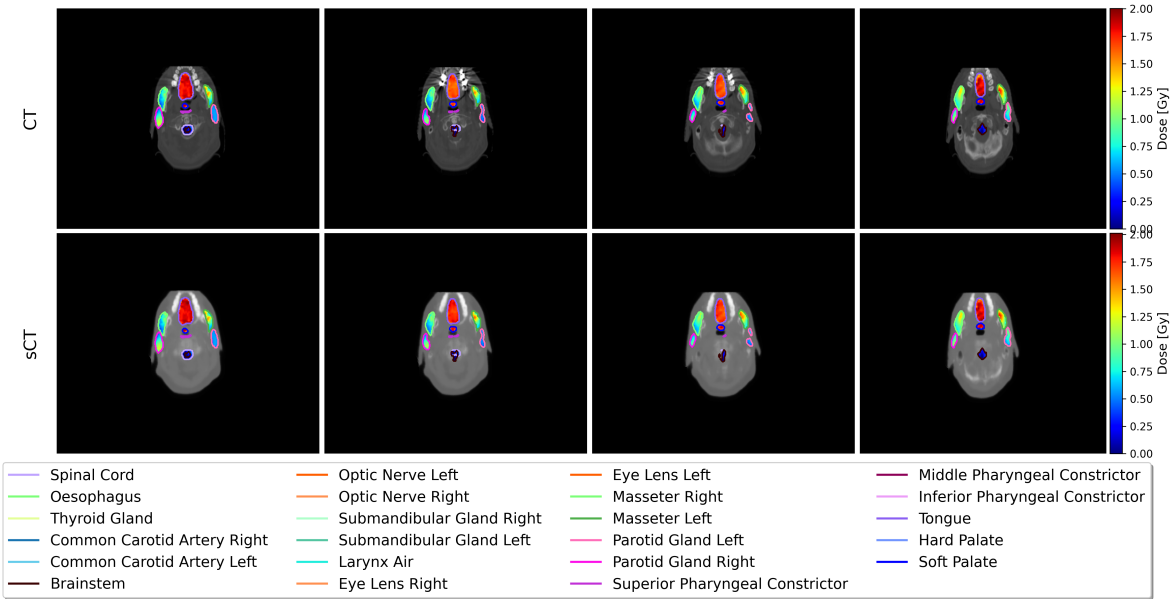


Figure 5.19: Slices 77, 78, 79 and 80 of the dose distribution on the CT and sCT for the best overall case (1HNA085) in the IMPT plan. The GTV are the tongue, hard and soft palates. The $\gamma_{2\%/2\text{mm}}$ pass rate is 99.78% and $MAE_{\text{target dose}} = 0.01$ Gy.

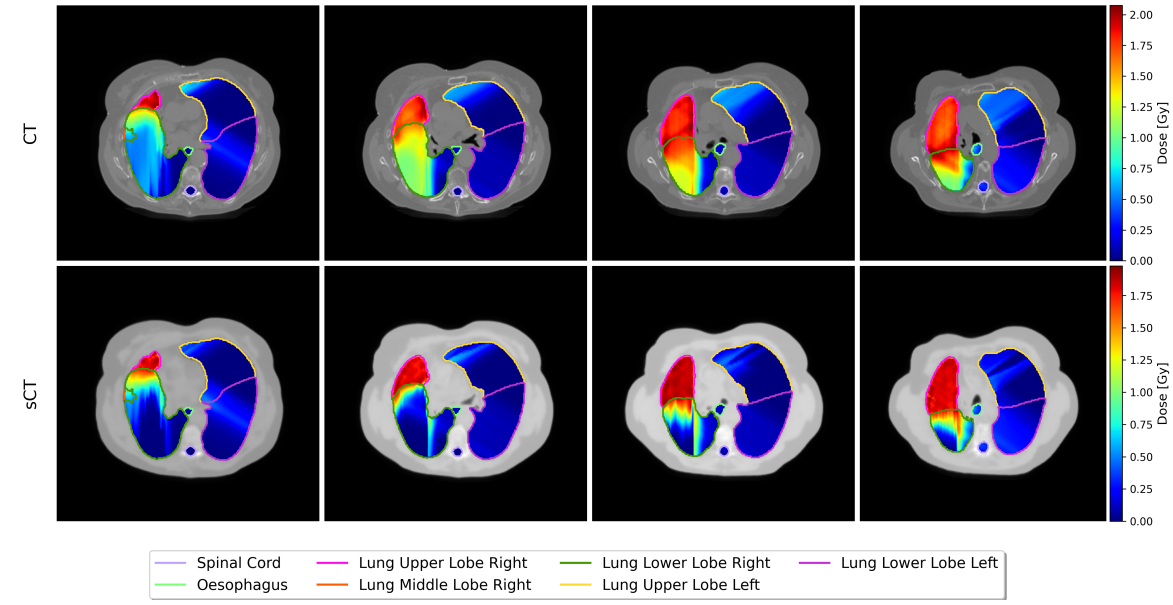


Figure 5.20: Slices 45, 50, 55, and 60 of the dose distribution on the CT and sCT for the IMPT plan for patient 1THA028. The GTV is the upper right lobe. The $\gamma_{2\%/2\text{mm}}$ pass rate is 65.41% and $MAE_{\text{target dose}} = 0.04$ Gy.

Chapter 6

Discussion

This chapter provides the study's main findings (section 6.1) and a thorough analysis of the results obtained and their comparison with the literature (section 6.2). More details regarding the constraints of the datasets (section 6.3), the study's limitations (section 6.4) and possible future research (section 6.5) will also be discussed in this chapter.

6.1 Main findings

The current study, which focused on the generation of CT-like images from MRI, found various important findings. Firstly, optimal HP encountered for the AE and cGAN 2D models, which were obtained through a cascade-type search, resulted in better AE validation and testing performance compared to the cGAN in both datasets. Secondly, the combination of the five trained models originated from cross-validation with the AE, and cGAN produced a more robust ensemble of models when compared to each of the five individual models. In this work, the combination or ensemble was calculated as the pixel-wise mean of the test predictions across the five models and presented better image quality metrics than an individual model. Third, the multi-region model (trained in data from all regions) tested in both single and mixed regions datasets, outperformed the models that were trained and tested in individual regions. Finally, dose calculations computed on treatment plans based on sCT were consistent with those based on CT.

6.2 Literature comparison

A comparison with relevant sCT methods in literature was made to further analyse the results obtained in this work. AE consistently outperforming cGAN is consistent with [70], the SynthRAD2023 final report, that concluded that teams who used 2D encoder-decoder models had superior results than those who employed GANs. Although the exact reasons for this are unclear, they may be related to the broader range of HP tuned for cGAN, which increases the complexity and hinders the achievement of an optimal configuration, since it is a time-consuming process. Another possible reason is the inherent difficulty of cGAN convergence, as the train-

ing is adversarial, leading to instability. Additionally, whether the cGAN's poorer performance is explained by those above-mentioned issues, dataset sizes, multi-centre image acquisition, or other factors of the end-to-end pipeline, remains inconclusive. Despite SynthRAD2023 test results showing statistically significant differences between the AE and cGAN only on the MAE (Figure 5.5), and SynthRAD2025 test results revealed that the null hypothesis could be rejected on PSNR, MAE and SSIM (Figure 5.6), the differences in metric values were marginal. Nonetheless, the AE achieved better performance in both tests. For non statistically significant differences, it seems possible that these were achieved due to the small sample size, which could have restricted the statistical power of the analysis, or the differences between the values being compared were minimal.

In contrast to what was done in validation, where metrics were computed as the average over all 2D normalized slices of the 3D volume (with the exception of the MAE, which was computed using masked 2D HU slices for immediate comparison with the literature), the testing metrics of the final model from each dataset were calculated using the 3D masked CT and sCT volumes in HU to follow best standard practices in a clinical research (Table 5.19). This adjustment in the MAE during the validation process was made to enable a comparison with other studies at the moment, as the MAE calculated on normalised images resulted in smaller values (e.g., MAE=0.09) while when calculated on images in HU, its values were larger (e.g., MAE=90 HU). For the MAE, masking was applied to focus the error computation in the region of interest, avoiding the influence of black pixels in the background areas that could either underestimate or overestimate its value. This masking was not applied to other metrics at first to assess the global perceptual quality of the generated images. Since the observational results revealed a monotonic trend, in which averaging the 2D slices metric values tended to be overestimated compared to the 3D masked volume calculations, the conclusions drawn remain valid. While in validation, the computed PSNR, SSIM and MS-SSIM were not assigned a fixed intensity range, for the 3D approach, a data range of 4000 HU (approximately: $3000 - (-1024)$) was defined to enable a fair comparison. A comparison of the mean MAE test results in the various regions, reveals that the pixel-wise differences were the highest in the HN set (149.40 ± 50.99 HU), due to images from centre C being acquired in a limited field of view, which made it nearly impossible for the model to generate the missing slices. Then, the TH set results showed a mean MAE of 118.46 ± 31.27 HU, possibly due to the presence of air in the lungs, while the best mean MAE value (106.99 ± 30.93 HU) was observed in the AB region, which may be attributed to having less air. The presence of air affects the results by introducing large intensity differences between CT and sCT, since the air has values around -1000 HU in CT images, and the input, MRI, does not capture air effectively (0 to no signal), resulting in the model struggling to generate air in the sCT.

As the SynthRAD2023 challenge paper, [70], reported PSNR, MAE, and SSIM metrics computed on 3D masked volumes in HU, the same approach was applied for testing the resultant model in this study (Table 5.19). The results revealed that the metrics obtained in different centres did not reject the null hypothesis for the pelvic data, which is consistent with the findings in [70]. Additionally, the mean SSIM results achieved in this study (0.84 ± 0.03) are in

line with the inter quartile range (across 19 teams) reported in the challenge (0.8 to 0.9). When considering MAE values, the present study findings were 71.69 ± 14.28 HU, which is lower than those achieved in average on the challenge (79.40 ± 28.30 HU). Nevertheless, the winning team outcomes averaged by the pelvic and brain data were 58.83 ± 13.41 HU (MAE), 29.61 ± 1.79 dB (PSNR), and 0.89 ± 0.03 (SSIM), surpassing the results of this work, possibly due to a superior hybrid 3D patch-based Convolutional Neural Network (CNN) and transformer U-Net network model approach.

In comparison to previous studies, SynthRAD2025 test results computed for 3D images in HU from HN region (Table 5.19), are inferior to those reported in [71], which employed a U-Net model with a four-channel input (from different MRI sequences - T1 + T1C + T1DixonC-water + T2). The results achieved were 71.31 ± 12.40 HU (MAE), 29.23 ± 1.29 dB (PSNR), and 0.85 ± 0.03 (SSIM), justifiably better due to the multi-modal input, providing more anatomical information. For AB and TH images generation, the study in [72] employed a U-Net to generate pCT from deformable CT-MRI pairs acquired for MR-guided radiotherapy. The findings in this study for 6 test images from AB region were 150 ± 55.4 HU (MAE), 22.9 ± 3.1 dB (PSNR), and 0.91 ± 0.01 (SSIM), and for 6 test images from TH region were 134 ± 27.8 HU (MAE), 23.0 ± 1.7 dB (PSNR), and 0.89 ± 0.01 (SSIM). Apart from the SSIM values reported in their work, which were superior when compared to SSIM from this study, the other two metrics showed inferior performance compared to those from this work. It seems possible that these results may have something to do with the number of images used to test the model, given that this study used a larger set available for testing, such as 16 images for the AB region and 18 for the TH region, significantly improving the model reproducibility. Regarding the mean MS-SSIM results for the HN dataset, the values obtained (0.80 ± 0.09) were superior to those reported by [73] (0.677 ± 0.019), who used the pix2pixHD neural network, an improved version of the original pix2pix, which is based on a cGAN architecture, to compare their results. This architecture is more similar to the AE used in this work. This difference is possibly attributed to their small dataset size of only 8 cases compared to 15 cases used in the present study.

Although the two editions of the SynthRAD challenge (SynthRAD2023 and SynthRAD2025) used the CT as reference for planning optimisation, meaning that the sCT plan was recalculated to each patient but not replanned, this work aimed to simulate a real MR-only workflow, where only a sCT would be available. Therefore, the treatment planning optimisation was performed solely on the sCT. Here, both image modalities used the segments from CT generated by the TotalSegmentator tool, as the geometric consistency metrics previously indicated that the sCT segments were of poor quality (Figure 5.10 and Figure 5.11). The geometric consistency metrics (mDICE and HD95) were not compared with the literature, as most studies doing research on the image-to-image translation task on MRI and CT images do not report these metrics. The study by Koivula et al. [74] showed that there were no significant changes in gamma index pass rates between IMPT plans that were first optimised in CT and then recalculated on heterogeneous sCT, and those that were first optimised on heterogeneous sCT and then recalculated on CT. Moreover, in the HN region, where tissue density is uniform and the cranium is an almost uniform spherical shell, plan optimisation in sCT achieved higher gamma pass rates. These were

affected by patient shape, but were less sensitive to density variations. In the present work, the highest gamma pass rate values were observed precisely for the HN region, with a mean value of $98.58 \pm 1.12\%$ under the 2%/2mm gamma index criteria, and a mean dose difference of 0.01 Gy or 1%. These results fall into those reported in [74], who achieved a gamma pass rate of $99.5\%(98.0 - 100)$ for the same criteria and mean dose difference at the target of 1.5%, which was slightly superior perhaps due to the localization of the tumour (brain in the paper vs. oral cavity in this work) and lower uncertainty in sCT generation. Conversely, the study found that in pelvic regions, where anatomical complexities increased, due to bones irregular shapes and heterogeneous composition of bone tissue, which affects dose modulation, the gamma pass rates tended to be lower for treatment plans optimised in sCT compared to those optimised in CT. This observation may justify the lower gamma pass rate value for the TH region in this work, as it is full of bone structures. To the best of our knowledge, no studies were found that computed dose calculations, such as dose differences, DVH or gamma pass rates on AB and TH images with the treatment plan optimised on the sCT and applied to the CT for re-calculation, without further optimisation.

When comparing centres' results of the $MAE_{\text{target dose}}$ metric in the AB region, values were slightly lower (better) when using images from centre B, where MRI sequences of the balanced steady-state free-precession type (T2/T1-weighted contrast) were acquired. This may be partly explained by the higher number of patients in this centre for the AB region compared to the other two centres. For both TH and HN regions, superior $MAE_{\text{target dose}}$ values were observed for centre A. Furthermore, the other two dose metrics reported better values within each region when using images from centre A, which can be justified by the larger number of images this centre provided for training the model across all regions. Taking this into account, and the fact that the MRI sequences used at this centre were variants of the T1-weighted gradient echo sequence, it can be concluded that the developed model performs better with MR images of this sequence type.

The outlier of this study regarding dose calculations is the 1THA028 case, which was optimised for the GTV (right upper lobe) and a portion of the OAR, since from the beam perspective these structures overlapped for the same gantry angles. However, upon verification, there was no real overlap between segments. Despite this, the DVH metric for the outlier was higher than the mean values in the same region (TH), indicating that the treatment plans in both CT and sCT were less comparable than other cases. The $MAE_{\text{target dose}}$ was lower than the average, which means the difference between the planned and delivered dose to the target was small, and thus, a high accuracy in delivering the dose to the target region. For the gamma pass rate, the value for the outlier was lower than the mean, which means that fewer points in the dose distribution passed the gamma evaluation criteria of 2%/2mm, thereby not achieving agreement across all dose points between planned and evaluated dose distributions, possibly caused by the optimisation involving structures that appeared to be overlapped.

Concerning the final generator architectures, both incorporated instance normalisation layers, which yielded superior results than batch normalisation, particularly more significant in the

cGAN case. These findings are in agreement with those obtained in [75], where replacing batch normalisation with instance normalisation in certain deep neural networks for image style transfer considerably improves the quality of generated images. The style transfer of this work is the transformation of an image from one modality to another in order to match the target modality, while preserving the anatomical content.

Moreover, the results of this study showed that the combination of the five trained models originated from CV with the AE produced a more robust ensemble of models. In this work, the combination or ensemble was calculated as the pixel-wise mean of the test predictions across the five models and presented better image quality metrics than an individual model, which supports the idea that the ensemble masked the overfitting of an individual model. This is in agreement with [76], which concluded that the ensemble of all networks through averaging yielded the best metric results, for models trained on images from one scanner with low-field strength and tested across MR scanners from different vendors and higher field strengths.

The multi-region model's superiority when compared to the individual region models can be justified with several explanations. Although it was initially expected that models trained on sets of single anatomical regions outperformed others, due to the model's attention on learning region-specific features from each region, this was not observed. A possible explanation is the limited number of images available to train and test each model, which may hinder generalisation. In contrast, the model trained in multiple regions revealed superior performance, which was trained with a larger and more diverse set. These results should be interpreted with caution because the dataset contained a mix of AB images in the TH set and vice versa, potentially augmenting the model's ability to learn these specific region features, thereby contributing to the improved testing performance on AB and TH sets and worsening the evaluation on HN sets.

6.3 Constraints of the datasets

In SynthRAD2023 ([55]), the mono-region and multi-centric dataset, differences in CT and MRI images may have been affected by physiological factors, such as bladder filling, air pockets in the rectum or bowel, and peristaltic motion due to time differences in acquisition. Furthermore, images provided by the challenge were not raw, because they were subjected to reconstruction parameters used in the clinical protocol.

SynthRAD2025 ([56]), which consists of multi-region pairs acquired in multi-centres, included various limitations that might have impacted the results. First of all, the position of the arms was different in certain MRI and CT acquisitions, and a restricted number of cases contained artefacts caused by metal implants. Each centre's definition of anatomical regions and imaging protocols varied, which integrated some thoracic images in the abdominal dataset and vice versa, adding a biased on the training and evaluation of individual models. Moreover, the patient outline, which accounts for a dilation margin, varies across datasets and patients due to patient variability upon definition of the body mask. Another limitation is related to the acquisition of MRI scans with a limited field of view in HN subset of centre C, which affected

the accurate rigid registration and hindered the sCT generation, leading to poorer image quality and DVH metrics. At last, subsets from AB and TH regions of centre B were cropped to exclude edge artifacts. However, some of those might still be included.

6.4 Limitations

The present study is not without limitations. The model fails to produce high-resolution images indistinguishable to the human eye, typical of AE outputs, which may compromise the generalisation of the model. This issue arises during compression, when the encoder reduces the input to a smaller latent representation that excludes high-frequency details, such as textures, which then cannot be fully reconstructed by the decoder. In this context, the AE model is fallible, particularly when the input (MRI) was acquired with a limited field of view, such as the HN images from centre C, resulting in a model incapable of producing the missing slices from the real CT. In fact, the worst-performing patient from the HN set belonged to centre C with a MAE value of 300.36 HU. Additionally, the generated images do not include certain anatomical structures, as proven by the missing segmentations identified with the TotalSegmentator tool and the geometric consistency metrics, depending on the anatomical region. For the existing segmentations, the geometric consistency metrics were very low, probably due to the model's difficulty in generating fine anatomical detail, specially bone. As a final limitation, the resultant model does not account for the heterogeneity of the datasets, which comprise data from various centres, vendors and regions, as well as the differences in imaging protocols and scanners within centres.

6.5 Future research

Future research directions should include developing models capable of addressing variability among multiple centres and vendors or exploring different architectures, such as the MC-DDPM, which is based on diffusion and outperforms traditional GANs and their variants in generating high-quality sCTs [3]. Thus, this would provide more insights into how model performance would be impacted, contributing to a more robust generation of sCT. Additionally, a critical caveat that needs special attention is the limited field of view in MRI acquisitions. A starting point to solve this could be to adapt the model architecture or evaluation methods to these specific cases.

Chapter 7

Conclusions

This study set out to explore how several general and anatomy-specific 2-dimensional machine-learning models, using AE and cGAN, can generate realistic synthetic-CT volumes from MR images, and assess their potential for MR-only radiotherapy planning with proton therapy. Initially, the models were trained and optimised in pelvic data, and further validation of their generalisability was performed using multi-regions data, including abdominal, thoracic and head and neck images. The results revealed that the 2D AE model consistently outperformed cGAN in both scenarios. Evaluation assessment of the final multi-regions model (model trained in all regions) was performed in three categories of metrics: image quality, geometric consistency and proton therapy dose calculations. Among all regions, the head and neck dosimetry results proved to be clinically relevant.

Furthermore, the established and achieved research goals suggest that the 2D AE models can produce sCT volumes useful for clinical applications, particularly for proton dose calculations, when using a larger and heterogeneous dataset in training and evaluation. While the overall sCT images quality was moderate, the dosimetry for proton therapy showed clinical feasibility for MR-only radiotherapy planning for the head and neck region, particularly when using T1-weighted MR images.

Moreover, the development of robust machine-learning models for image-to-image translation reinforces the potential of MR-only radiotherapy planning to become a more reliable and broader clinical reality, improving patient safety and reducing the cancer burden at the treatment planning stage. Future developments could focus on improving sCT images resolution by designing more high-level architectures and employing an anatomically conditioned approach, rather than using a single architecture type and fixed HP for all regions. Additionally, greater attention to MR images acquired in a limited field of view could enhance the realism of sCTs and contribute to the development of more robust and reliable neural network models.

Bibliography

[1] WHO, "Cancer - detailed fact sheets." https://www.who.int/news-room/fact-sheets/detail/cancer, 2025.

[2] Y. Lievens, J. M. Borras, C. Grau, and A. Aggarwal, "Value-based radiotherapy: A new chapter of the estro-hero project," *Radiotherapy and Oncology*, vol. 160, pp. 236–239, 7 2021.

[3] M. A. Bahloul, S. Jabeen, S. Benoumhani, H. A. Alsaleh, Z. Belkhatir, and A. Al-Wabil, "Advancements in synthetic CT generation from MRI: A review of techniques, and trends in radiation therapy planning," *Journal of applied clinical medical physics*, vol. 25, 11 2024.

[4] M. Boulanger, J. C. Nunes, H. Chourak, A. Largent, S. Tahri, O. Acosta, R. D. Crevoisier, C. Lafond, and A. Barateau, "Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review," *Physica Medica: European Journal of Medical Physics*, vol. 89, pp. 265–281, 9 2021.

[5] S. Tahri, A. Barateau, C. Cadin, H. Chourak, S. Ribault, F. Nozahic, O. Acosta, J. A. Dowling, P. B. Greer, A. Largent, C. Lafond, R. D. Crevoisier, and J. C. Nunes, "A high-performance method of deep learning for prostate MR-only radiotherapy planning using an optimized Pix2Pix architecture," *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB)*, vol. 103, pp. 108–118, 11 2022.

[6] J. A. Dowling, J. Sun, P. Pichler, D. Rivest-Hénault, S. Ghose, H. Richardson, C. Wratten, J. Martin, J. Arm, L. Best, S. S. Chandra, J. Fripp, F. W. Menk, and P. B. Greer, "Automatic substitute computed tomography generation and contouring for magnetic resonance imaging (MRI)-alone external beam radiation therapy from standard MRI sequences," *International Journal of Radiation Oncology*Biology*Physics*, vol. 93, pp. 1144–1153, 12 2015.

[7] W. R. Crum, T. Hartkens, and D. L. Hill, "Non-rigid image registration: theory and practice," *The British journal of radiology*, vol. 77 Spec No 2, 2004.

[8] A. Largent, A. Barateau, J. C. Nunes, C. Lafond, P. B. Greer, J. A. Dowling, H. Saint-Jalmes, O. Acosta, and R. de Crevoisier, "Pseudo-CT generation for MRI-only radiation therapy treatment planning: Comparison among patch-based, atlas-based, and bulk density

methods," *International Journal of Radiation Oncology*Biology*Physics*, vol. 103, pp. 479–490, 2 2019.

[9] A. Bahrami, A. Karimian, and H. Arabi, "Comparison of different deep learning architectures for synthetic CT generation from MR images," *Physica Medica*, vol. 90, pp. 99–107, 10 2021.

[10] L. Taylor Francis Group, "(pdf) handbook of radiotherapy physics: Theory and practice."

[11] R. Baskar, K. A. Lee, R. Yeo, K.-W. Yeoh, R. Baskar, and M. Phil, "Cancer and radiation therapy: Current advances and future directions," *Int. J. Med. Sci*, vol. 9, pp. 193–199, 2012.

[12] N. Mail, K. M. Alshamrani, R. N. Lodhi, E. Khawandanh, A. Saleem, B. Khan, M. Alghamdi, M. Nadershah, M. S. Althaqafy, A. Subahi, and S. M. Alghamdi, "Evaluation of positioning accuracy in head-and-neck cancer treatment: A cone beam computed tomography assessment of three immobilization devices with volumetric modulated arc therapy," *Journal of biological methods*, vol. 11, p. e99010025, 2024.

[13] H. I. Pass, D. Ball, G. V. Scagliotti, and IASLC, eds., *IASLC Thoracic Oncology*. Elsevier, 2 ed., 2018.

[14] S. J. Gardner, J. Kim, and I. J. Chetty, "Modern radiation therapy planning and delivery," *Hematology/oncology clinics of North America*, vol. 33, pp. 947–962, 12 2019.

[15] S. Devic, "MRI simulation for radiotherapy treatment planning," *Medical Physics*, vol. 39, pp. 6701–6711, 2012.

[16] M. T. Spiotto and P. P. Connell, "Strategies to overcome late complications from radiotherapy for childhood head and neck cancers," *Oral and maxillofacial surgery clinics of North America*, vol. 28, pp. 115–126, 2 2016.

[17] N. G. Burnet, S. J. Thomas, K. E. Burton, and S. J. Jefferies, "Defining the tumour and target volumes for radiotherapy," *Cancer Imaging*, vol. 4, p. 153, 2004.

[18] K. K. Brock, S. Mutic, T. R. McNutt, H. Li, and M. L. Kessler, "Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the aapm radiation therapy committee task group no. 132: Report," *Medical Physics*, vol. 44, pp. e43–e76, 7 2017.

[19] S. Dragojevic, J. Ji, P. K. Singh, M. A. Connors, R. W. Mutter, S. C. Lester, S. M. Talele, W. Zhang, B. L. Carlson, N. B. Remmes, S. S. Park, W. F. Elmquist, S. Krishnan, E. J. Tryggestad, and J. N. Sarkaria, "Preclinical risk evaluation of normal tissue injury with novel radiosensitizers," *International Journal of Radiation Oncology*Biology*Physics*, vol. 111, pp. e54–e62, 12 2021.

[20] R. Holla, "The physics of radiotherapy x-rays and electrons," *Journal of Medical Physics*, vol. 49, p. 487, 7 2024.

[21] H. M. Adil, S. F. Abbas, S. awad kadhim, A. H. Mohmmed, and A. H. Sharad, "Radiation therapy systems using linear accelerator, importance of its components and their development," *GSC Advanced Research and Reviews*, vol. 20, pp. 022–029, 9 2024.

[22] K. Lapen and Y. Yamada, "The development of modern radiation therapy," *Current Physical Medicine and Rehabilitation Reports*, vol. 11, pp. 131–138, 6 2023.

[23] H. Liu and J. Y. Chang, "Proton therapy in clinical practice," *Chinese Journal of Cancer*, vol. 30, p. 315, 2011.

[24] A. C. Moreno, S. J. Frank, A. S. Garden, D. I. Rosenthal, C. D. Fuller, G. B. Gunn, J. P. Reddy, W. H. Morrison, T. D. Williamson, E. B. Holliday, J. Phan, and P. Blanchard, "Intensity modulated proton therapy (IMPT) – the future of IMRT for head and neck cancer," *Oral Oncology*, vol. 88, pp. 66–74, 1 2019.

[25] J. P. Gibbons, "Khan's the physics of radiation therapy," *Journal of Medical Physics*, vol. 45, p. 134, 8 2020.

[26] J. A. Purdy, "Intensity-modulated radiotherapy: Current status and issues of interest," *International Journal of Radiation Oncology Biology Physics*, vol. 51, pp. 880–914, 11 2001.

[27] L. M. Pudsey, D. Cutajar, A. Wallace, A. Saba, L. Schmidt, A. Bece, C. Clark, Y. Yamada, G. Biasi, A. Rosenfeld, and J. Poder, "The use of collimator angle optimization and jaw tracking for vmat-based single-isocenter multiple-target stereotactic radiosurgery for up to six targets in the varian eclipse treatment planning system," *Journal of Applied Clinical Medical Physics*, vol. 22, pp. 171–182, 9 2021.

[28] T. Pavel, "Feasibility of magnetic resonance imaging-based radiation therapy for brain tumour treatment," 2018. Master's thesis, Aalto University. Available at `https://aaltodoc.aalto.fi/server/api/core/bitstreams/196825e1-8d24-4083-a182-8339a62cd78b/content`.

[29] P. J. Hoskin and V. Goh, "Radiotherapy in practice – imaging," *The British Journal of Radiology*, vol. 83, p. 993, 2010.

[30] F. Esmaeili, M. Johari, P. Haddadi, and M. Vatankhah, "Beam hardening artifacts: Comparison between two cone beam computed tomography scanners," *Journal of Dental Research, Dental Clinics, Dental Prospects*, vol. 6, p. 49, 2012.

[31] M. Mahesh, "The essential physics of medical imaging, third edition," *Medical physics*, vol. 40, 7 2013.

[32] C. Glide-Hurst, D. Chen, H. Zhong, and I. J. Chetty, "Changes realized from extended bit-depth and metal artifact reduction in CT," *Medical Physics*, vol. 40, p. 061711, 6 2013.

[33] H. P. Wieser, E. Cisternas, N. Wahl, S. Ulrich, A. Stadler, H. Mescher, L. R. Muller, T. Klinge, H. Gabrys, L. Burigo, A. Mairani, S. Ecker, B. Ackermann, M. Ellerbrock, K. Parodi, O. Jakel, and M. Bangert, "Development of the open-source dose calculation and optimization toolkit matrad," *Medical Physics*, vol. 44, pp. 2556–2568, 6 2017.

[34] U. Schneider, E. Pedroni, and A. Lomax, "The calibration of CT hounsfield units for radiotherapy treatment planning," *Physics in Medicine and Biology*, vol. 41, pp. 111–124, 1 1996.

[35] R. Bitar, G. Leung, R. Perng, S. Tadros, A. R. Moody, J. Sarrazin, C. McGregor, M. Christakis, S. Symons, A. Nelson, and T. P. Roberts, "MR pulse sequences: What every radiologist wants to know but is afraid to ask," *Radiographics*, vol. 26, pp. 513–537, 3 2006.

[36] S. Mastrogiacomo, W. Dou, J. A. Jansen, and X. F. Walboomers, "Magnetic resonance imaging of hard tissues and hard tissue engineered bio-substitutes," *Molecular Imaging and Biology*, vol. 21, pp. 1003–1019, 12 2019.

[37] A. M. Owrangi, P. B. Greer, and C. K. Glide-Hurst, "MRI-only treatment planning: Benefits and challenges," *Physics in Medicine and Biology*, vol. 63, 2 2018.

[38] R. K. Mishra, G. Y. Reddy, and H. Pathak, "The understanding of deep learning: A comprehensive review," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[39] M. Tahmid, M. S. Alam, N. Rao, and K. M. A. Ashrafi, "Image-to-image translation with conditional adversarial networks," *Proceedings of 2023 IEEE 9th International Women in Engineering (WIE) Conference on Electrical and Computer Engineering, WIECON-ECE 2023*, pp. 468–472, 11 2016.

[40] "Introduction to machine learning: Generative adversarial networks (GANs)." https://developers.google.com/machine-learning/gan/, 2024.

[41] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 12 2016. Preprint on webpage at `https://arxiv.org/abs/1701.00160v4`.

[42] D. Saxena and J. Cao, "Generative adversarial networks (GANs survey): Challenges, solutions, and future directions," 4 2020. Preprint on webpage at `https://arxiv.org/abs/2005.00065v4`.

[43] U. Michelucci, "An introduction to autoencoders," 1 2022. Preprint on webpage at `https://arxiv.org/pdf/2201.03898`.

[44] J. Chen, G. Yang, X. Zhang, J. Peng, T. Zhang, J. Zhang, J. Han, and V. Grau, "Unsupervised Patch-GAN with targeted patch ranking for fine-grained novelty detection in medical imaging," 1 2025. Preprint on webpage at `https://arxiv.org/pdf/2501.17906`.

[45] A. Largent, A. Barateau, J. C. Nunes, E. Mylona, J. Castelli, C. Lafond, P. B. Greer, J. A. Dowling, J. Baxter, H. Saint-Jalmes, O. Acosta, and R. de Crevoisier, "Comparison of deep learning-based and patch-based methods for pseudo-CT generation in MRI-based prostate dose planning," *International journal of radiation oncology, biology, physics*, vol. 105, pp. 1137–1150, 12 2019.

[46] B. Texier, C. Hémon, A. Queffélec, J. Dowling, I. Bessieres, P. Greer, O. Acosta, A. Boue-Rafle, R. de Crevoisier, C. Lafond, J. Castelli, A. Barateau, and J. C. Nunes, "3d unsupervised deep learning method for magnetic resonance imaging-to-computed tomography

synthesis in prostate radiotherapy," *Physics and imaging in radiation oncology*, vol. 31, 7 2024.

[47] S. Pan, E. Abouei, J. Wynne, T. Wang, R. L. J. Qiu, Y. Li, C.-W. Chang, J. Peng, J. Roper, P. Patel, D. S. Yu, H. Mao, and X. Yang, "Synthetic CT generation from MRI using 3d transformer-based denoising diffusion model," 5 2023. Preprint on webpage at `https://arxiv.org/abs/2305.19467v1`.

[48] S.-H. Tsang, "Review: V-Net — volumetric convolution (biomedical image segmentation)," 2019. Available at `https://towardsdatascience.com/review-v-net-volumetric-convolution-biomedical-image-segmentation-aa15dbaea974`.

[49] P. Ruiz, "Understanding and visualizing ResNets," 2021. Available at `https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8`.

[50] H. Chen, "Challenges and corresponding solutions of generative adversarial networks (GANs): A survey study," *Journal of Physics: Conference Series*, vol. 1827, p. 012066, 3 2021.

[51] Medium, "Diffusion models: A comprehensive high-level understanding." Available at `https://medium.com/@researchgraph/diffusion-model-comprehensive-high-level-understanding-55d6ecad2cba`.

[52] J. Terven, D. M. Cordova-Esparza, A. Ramirez-Pedraza, E. A. Chavez-Urbiola, and J. A. Romero-Gonzalez, "Loss functions and metrics in deep learning," 7 2023.

[53] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1398–1402, 2003.

[54] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. Pluim, "Elastix: A toolbox for intensity-based medical image registration," *IEEE Transactions on Medical Imaging*, vol. 29, pp. 196–205, 1 2010.

[55] A. Thummerer, E. van der Bijl, A. J. Galapon, J. J. Verhoeff, J. A. Langendijk, S. Both, Cornelis, A. van den Berg, and M. Maspero, "Synthrad2023 grand challenge dataset: generating synthetic ct for radiotherapy," *Medical Physics*, vol. 50, pp. 4664–4674, 3 2023.

[56] A. Thummerer, E. van der Bijl, A. J. Galapon, F. Kamp, M. Savenije, C. Muijs, S. Aluwini, R. J. H. M. Steenbakkers, S. Beuel, M. P. W. Intven, J. A. Langendijk, S. Both, S. Corradini, V. Rogowski, M. Terpstra, N. Wahl, C. Kurz, G. Landry, and M. Maspero, "SynthRAD2025 grand challenge dataset: generating synthetic cts for radiotherapy," 2 2025. Preprint on webpage at `https://arxiv.org/pdf/2502.17609`.

[57] E. Goceri, "Medical image data augmentation: techniques, comparisons and interpretations," *Artificial Intelligence Review*, vol. 56, pp. 12561–12605, 11 2023.

[58] MONAI, "Randflip transform — monai 1.5.0 documentation." Available at `https://docs.monai.io/en/stable/transforms.html#randflip`.

[59] MONAI, "Rand2delasticd transform — monai 1.5.0 documentation." Available at `https://docs.monai.io/en/stable/transforms.html#rand2delastic`.

[60] MONAI, "Randrotate90d transform — monai 1.5.0 documentation." Available at `https://docs.monai.io/en/stable/transforms.html#randrotate90d`.

[61] W. Ullah, S. Ilyas, H. Naveed, and S. Ali, "An integrated approach to image quality: comparative analysis of bilinear and nearest neighbor interpolation," *Big Data and Computing Visions*, vol. 5, pp. 24–36, 3 2025.

[62] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," *Advances in Neural Information Processing Systems*, pp. 2234–2242, 6 2016.

[63] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "Radimagenet: An open radiologic deep learning research dataset for effective transfer learning," *Radiology: Artificial Intelligence*, vol. 4, 9 2022.

[64] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "TotalSegmentator: robust segmentation of 104 anatomical structures in CT images," *Radiology: Artificial Intelligence*, vol. 5, 8 2022.

[65] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, pp. 203–211, 2 2021.

[66] S. G. Challenge, "Metrics & ranking tab." Available at https://synthrad2025.grand-challenge.org/metrics-ranking/.

[67] J. Ferlay, M. Ervik, F. Lam, M. Laversanne, M. Colombet, L. Mery, M. Piñeros, A. Znaor, I. Soerjomataram, and F. Bray, "Global cancer observatory: Cancer today." `https://gco.iarc.who.int/today`, 2024.

[68] Y. Nilssen, O. T. Brustugun, L. Fjellbirkeland, Åslaug Helland, B. Møller, S. G. F. Wahl, and S. Solberg, "Distribution and characteristics of malignant tumours by lung lobe," *BMC Pulmonary Medicine*, vol. 24, pp. 1–10, 12 2024.

[69] O. M. Physics, "IMRT quality assurance." Available at https://oncologymedicalphysics.com/imrt-quality-assurance/.

[70] E. M. Huijben, M. L. Terpstra, A. J. Galapon, S. Pai, A. Thummerer, P. Koopmans, M. Afonso, M. van Eijnatten, O. Gurney-Champion, Z. Chen, Y. Zhang, K. Zheng, C. Li, H. Pang, C. Ye, R. Wang, T. Song, F. Fan, J. Qiu, Y. Huang, J. Ha, J. S. Park, A. Alain-Beaudoin, S. Bériault, P. Yu, H. Guo, Z. Huang, G. Li, X. Zhang, Y. Fan, H. Liu, B. Xin, A. Nicolson, L. Zhong, Z. Deng, G. Müller-Franzes, F. Khader, X. Li, Y. Zhang, C. Hémon, V. Boussot, Z. Zhang, L. Wang, L. Bai, S. Wang, D. Mus, B. Kooiman, C. A. Sargeant, E. G.

Henderson, S. Kondo, S. Kasai, R. Karimzadeh, B. Ibragimov, T. Helfer, J. Dafflon, Z. Chen, E. Wang, Z. Perko, and M. Maspero, "Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report," *Medical Image Analysis*, vol. 97, p. 103276, 10 2024.

[71] M. Qi, Y. Li, A. Wu, Q. Jia, B. Li, W. Sun, Z. Dai, X. Lu, L. Zhou, X. Deng, and T. Song, "Multi-sequence MR image-based synthetic CT generation using a generative adversarial network for head and neck MRI-only radiotherapy," *Medical Physics*, vol. 47, pp. 1880–1894, 4 2020.

[72] S. K. Kang, H. J. An, H. Jin, J. in Kim, E. K. Chie, J. M. Park, and J. S. Lee, "Synthetic CT generation from weakly paired MR images using cycle-consistent GAN for MR-guided radiotherapy," *Biomedical Engineering Letters*, vol. 11, pp. 263–271, 2021.

[73] Y. Li, S. Xu, Y. Lu, and Z. Qi, "CT synthesis from MRI with an improved multi-scale learning network," *Frontiers in Physics*, vol. 11, 1 2023.

[74] L. Koivula, L. Wee, and J. Korhonen, "Feasibility of MRI-only treatment planning for proton therapy in brain and prostate cancers: Dose calculation accuracy in substitute CT images," *Medical Physics*, vol. 43, pp. 4634–4642, 8 2016.

[75] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization." Preprint on webpage at `https://arxiv.org/abs/1701.02096.`

[76] L. Fetty, T. Löfstedt, G. Heilemann, H. Furtado, N. Nesvacil, T. Nyholm, D. Georg, and P. Kuess, "Investigating conditional gan performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion," *Physics in Medicine & Biology*, vol. 65, p. 105004, 5 2020.