A Study of Synthesis of CT Images from MRI Data with Guided Segmentation Masks

Xiang Chen Zhu

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

December 2025

© Xiang Chen Zhu, 2026

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Xiang Chen Zhu**

Entitled: **A Study of Synthesis of CT Images from MRI Data with Guided Segmentation Masks**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Yang Wang

_____ Examiner
Dr. Yiming Xiao

_____ Supervisor
Dr. Thomas Fevens

Approved by _____
Joey Paquet, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

A Study of Synthesis of CT Images from MRI Data with Guided Segmentation Masks

Xiang Chen Zhu

Magnetic Resonance Imaging (MRI) offers excellent soft-tissue contrast without ionizing radiation but lacks the quantitative attenuation information needed for applications such as radiotherapy planning and PET/MRI attenuation correction. This thesis introduces MAGNet, a mask-guided MR-to-CT synthesis framework that incorporates anatomy-aware priors into a Generative Adversarial Network (GAN) to generate high-fidelity synthetic CT (sCT) from pelvic MRI.

MAGNet uses segmentation masks automatically produced by TotalSegmentator to divide the translation into anatomically informed sub-tasks. Two specialized branches, conditioned on bone and soft-tissue masks, allow the network to focus on thin cortical structures, trabecular detail, and subtle density variations while preserving soft-tissue realism and suppressing artifacts. Outputs from these branches are adaptively fused through a learned blending mechanism, enhancing fidelity at bone–soft-tissue interfaces and in regions prone to motion or susceptibility distortions.

The framework is evaluated on a public pelvic MRI–CT dataset and a multi-scanner internal cohort, and benchmarked against representative learning-based baselines. MAGNet achieves consistently higher PSNR and SSIM, lower mean absolute error, and superior bone and soft-tissue reconstruction quality.

Requiring no manual contouring and adding minimal preprocessing, MAGNet is suited for integration into clinical workflows. By improving the structural fidelity and robustness of MR-to-CT translation, it advances MRI-only imaging pipelines toward reliable, radiation-free substitutes compatible with downstream planning, image-guided interventions, and quantitative analysis.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Thomas Fevens, for his invaluable guidance, patience, and continuous support throughout my Master's study. His mentorship was instrumental in shaping the direction of this research and navigating the challenges of academic inquiry.

I am also sincerely grateful to the members of my examination committee, Dr. Yang Wang and Dr. Yiming Xiao. I thank them for their time, their insightful comments, and their constructive feedback, all of which contributed significantly to the improvement of this thesis.

I extend a special note of appreciation to Professor Dr. Susanne Mayer at the Ludwig Maximilian University (LMU) of Munich. I am deeply thankful for the arrangement of my lab visit and for granting access to the large-scale internal clinical datasets that were essential for the validation of the MAGnet framework.

My thanks also go to my colleagues for their technical collaboration. I am grateful to Giles Michael Cheers, PhD candidate at LMU Munich, for his generous coding support and technical assistance. I also acknowledge Aloys Portafaix, PhD candidate at Polytechnique Montréal, for his valuable contributions to the preliminary groundwork of this project.

Most importantly, special thanks are owed to all the patients who consented to participate in this study. Their contribution is the foundation of medical research, and without them, advancements in clinical care would not be possible.

Finally, I would like to thank my family for their unwavering love and encouragement. To my fiancée, thank you for your endless support, patience, and understanding during the long hours of research and writing.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Cancer remains a leading cause of death worldwide, and radiation therapy is a cornerstone of its treatment, utilized in approximately 50% of all cancer cases [5]. The primary goal of modern external beam radiotherapy is to deliver a lethal dose of radiation to the tumor volume while sparing the surrounding healthy organs at risk (OARs) [6]. Achieving this balance requires precise treatment planning, which is fundamentally dependent on high-quality medical imaging.

Currently, the clinical standard for radiotherapy planning relies on a multi-modality workflow [7]. Computed Tomography (CT) serves as the primary imaging modality because it provides geometrically faithful information regarding tissue electron density. This electron density information is critical for accurate dose calculation, as the attenuation of therapeutic X-rays is linearly related to the electron density of the tissue it traverses [8]. Without this quantitative information, the treatment planning system cannot accurately model how radiation will be absorbed by the patient's body.

However, while CT is excellent for dose calculation and visualizing bony anatomy, it suffers from poor soft-tissue contrast. This limitation makes it difficult to precisely delineate tumor boundaries and adjacent soft-tissue OARs, particularly in complex anatomical regions such as the brain, head and neck, and pelvis [9]. To address this, Magnetic Resonance Imaging (MRI) is routinely acquired as a secondary modality. MRI offers superior soft-tissue contrast and can provide functional information, allowing for more precise target definition [10].

1.2 Motivation

Despite the benefits of using both CT and MRI, this dual-modality workflow introduces significant clinical and technical challenges. The integration of MRI into a CT-based workflow requires a process of aligning the MRI coordinate system with the CT coordinate system, which is called image registration. Even with advanced algorithms, registration is prone to systematic errors, often introducing geometrical uncertainties which could propagate through the entire treatment chain, potentially leading to geographical misses of the tumor or over-dosage of healthy tissue [11]. Furthermore, the acquisition of a planning CT exposes patients to additional ionizing radiation, a concern that becomes increasingly relevant for pediatric patients [12] or patients requiring multiple adaptive re-planning scans.

These limitations have motivated a paradigm shift toward MRI-only radiotherapy, where the CT scan is eliminated from the workflow entirely [13]. In this streamlined workflow, MRI is used for both target delineation and dose planning. However, the implementation of MRI-only radiotherapy faces a fundamental physical obstacle: MRI signal intensity correlates with proton density and tissue relaxation properties (T1, T2), not electron density. Consequently, MRI data cannot be directly used by standard treatment planning systems for dose calculation [14]. To bridge this gap, desired methods must be able to acquire image volumes that possess both the geometric fidelity of the MRI and the electron density information of a CT. A potential solution is to generate medical images across modalities, such as creating synthetic CT (sCT) images from MRI [15].

1.3 Challenges and limitations

The generation of high-quality sCTs is a challenging image-to-image translation problem. Some early methods include voxel-based approaches and multi-atlas based methods [16]. Voxel-based approaches require a time-consuming process of manual segmentation structures to predict the Hounsfield units (HU) from MRI. Multi-atlas methods depend on paired MRI and CT atlases to create deformable registration for sCT synthesis, therefore, the performance is severely constrained by the similarities within the pairs.

Deep Learning (DL), specifically Generative Adversarial Networks (GANs), has recently emerged

as the state-of-the-art solution for sCT generation. Unsupervised models such as CycleGAN [17, 18] have demonstrated the ability to produce synthetic CT images from unpaired CT images, essentially lifting the constraints of paired datasets. However, because the MR and sCT domains are not directly paired, the original CycleGAN has no built-in mechanism to enforce structural alignment during MR-to-CT synthesis, hence leading to inaccurate mappings and unconventional anatomy. One potential solution suggested by Phan et al [19] is to incorporate an extracted coarse mask to maintain shape consistency.

Still, a critical limitation remains in current DL frameworks. They typically treat the entire anatomical volume uniformly. This is particularly problematic in complex anatomical regions like the pelvis. In pelvic MRI, cortical bone usually has low signal intensity, which can easily be confused with air or bowel gas, since they all appear to be dark. When a standard GAN attempts to learn a mapping for the entire image at once, it often struggles to balance the distinct statistical properties of bone and soft tissue, leading to over-smoothed bone or hallucinated textures [3]. Therefore, there is an urgent need for an sCT generation framework that is anatomy-aware, which can disentangle the generation of rigid bony structures from the texture rich generation of soft tissues, without requiring labor intensive manual segmentation.

1.4 Contributions

To address the limitations of uniform synthesis models, this thesis proposes MAGnet, a dual-branch GAN for synthesizing CT from MRI in anatomically complex regions. In contrast to voxel driven approaches, MAGNet removes the dependency on manual segmentation by substituting it with a learning-based strategy. The specific contributions of this work are as follows:

1. We introduce a novel GAN architecture that splits the synthesis task into two parallel streams: a structure branch dedicated to accurate bone geometry reconstruction and a texture branch dedicated to soft-tissue fidelity. This design mitigates the trade-offs inherent in uniform translation models.

2. We present a strategy to leverage TotalSegmentator [4, 20], an automated segmentation tool, to derive categorized anatomical masks. This allows the model to be anatomy conditioned without

requiring manual contouring by clinicians, preserving the efficiency of the MRI-only workflow.

3. We validate the MAGnet framework on both a public dataset (Gold Atlas) and a large internal clinical cohort. We demonstrate that our approach yields statistically significant improvements in image quality metrics (MAE, PSNR, SSIM) compared to standard Pix2Pix [21] and CycleGAN baselines, and multiple advanced GAN variants, particularly in the challenging pelvic region.

1.5 Thesis Outline

The remainder of this thesis is structured as follows. In Chapter 2 , we present a comprehensive review of CT and MRI physics, surveys the evolution of clinical sCT generation methods, and summarizes the various strategies that have been proposed to address this problem.

In Chapter 3, we present the methodology of our proposed MAGnet framework. We begin by introducing the prior-derivation strategy and describing the dual-path generator architecture. We then detail the specialized loss functions used during training and explain how they interact to guide the learning process.

In Chapter 4, we describe the experimental setup, including the datasets and implementation details. We then report our comprehensive quantitative and qualitative evaluations, comparing the performance of MAGnet against standard baseline approaches.

In Chapter 5, we conclude by summarizing the key findings of this work, discussing its current limitations, and offering insights into potential directions for future research in MRI-only radiotherapy.

Chapter 2

Background

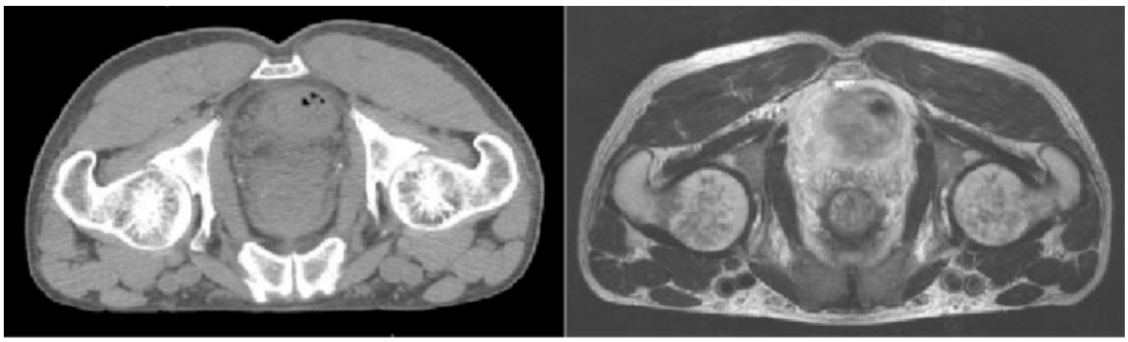
This chapter establishes the theoretical and clinical backgrounds which led to the inspiration for the thesis. As such, we start by introducing general modalities of medical imaging. Then we outline the evolution of sCT generation strategies, with a focus on the transition toward deep learning solutions.

2.1 Medical Imaging Modalities

2.1.1 Computed tomography

Computed tomography is an X-ray tomographic technique that provides cross-sectional images of the body through the use of X-ray measurements taken from different angles. Clinical CT is currently based on helical CT scanners, in which images are acquired using a continuously rotating X-ray tube and by moving the table on which the patient lies through the scan plane. The reconstruction algorithms, typically filtered back-projection (FBP) or iterative reconstruction techniques [22], generate a volumetric map of these attenuation coefficients. To standardize these values across different scanners and energy spectra, they are converted into Hounsfield Units (HU), defined as:

$$HU = 1000 \times \frac{\mu_{tissue} - \mu_{water}}{\mu_{water}} \tag{1}$$



(a)

Figure 2.1: CT (left) and MRI (right) identifying some of the bony pelvis anatomical landmarks on ischial spine. [1]

where μ represents the linear attenuation coefficient of the corresponding material. In CT scans, the Hounsfield scale is represented using greyscale intensity. Structures with higher X-ray attenuation show positive HU values and appear bright, whereas those with lower attenuation show negative values and appear dark. [23]

Overall, CT imaging provides highly accurate spatial detail, quantitative tissue-density information essential for dosimetry, and clear visualization of bone structures for detecting bone tumours. Its main limitations are the relatively poor soft-tissue contrast and the exposure of healthy tissue to ionizing radiation.

2.1.2 Magnetic Resonance Imaging

MRI is a non-invasive medical imaging modality that is used to generate three-dimensional detailed anatomical images of organs. It measures the signal emitted by protons in the body's water and fat molecules relaxing in a strong magnetic field after excitation by radiofrequency pulses.[24] Compared to CT, MRI provides excellent soft tissue contrast, which makes it the choice for screening pelvis, brain, spinal cord, and some head and neck tumours. [25] The primary limitation of MRI for radiotherapy is that, whereas CT allows electron density essential for dose calculations to be estimated from HU, MRI signal intensity does not correlate with radiation beam attenuation. As a result, patients are still exposed to extra radiation, since CT imaging is still required for treatment

planning.

2.2 Historical Evolution of Synthetic CT Generation

In order to use MRI alone for radiation therapy planning, it would require deriving HU directly from the MRI scans. Numerous techniques have been developed to generate MRI-based synthetic CT images, which can be broadly classified into four major categories: bulk density override, voxel-based approaches, atlas-based approaches, and learning-based approaches. The first three are considered more traditional techniques, whereas the last represent more recent developments. [26]

2.2.1 Bulk Density

The earliest methods developed for generating synthetic CT images from MRI scans relied on applying bulk density overrides for dose calculation. The simplest version of this approach assigns the entire patient volume a water-equivalent electron density. While this technique was tested in the brain [27] and prostate [28], it produced clinically unacceptable dose calculations compared to those based on the original heterogeneous densities. The average discrepancies across the whole volume exceeded 2%, especially when the beam passed through air cavities [29]. Furthermore, assigning a single density to all bone neglects the significant variation between spongy and cortical bone, which is critical for accurate dosimetry in regions where beams often traverse thick bony structure.

2.2.2 Voxel-based

Voxel-based approaches for sCT generation aim to predict HU values directly from MRI signal intensities, often by applying statistical regression models trained on multiple MRI sequences. When only conventional MRI sequences are available [30], manual segmentation of structures such as bone and airways is frequently required to distinguish tissues with similar signal characteristics. In some voxel-based implementations, individual voxel intensities are used independently to estimate HU values [31], which means spatial information such as neighborhood context is ignored. This simplification often leads to artifacts and inaccuracies at air/soft-tissue and bone/soft-tissue interfaces. To address this issue, Johansson et al. [32] introduced spatial features including the

voxel's (x, y, z) coordinates and distance to the body contour into the regression model. This addition significantly improved the accuracy and visual quality of sCT generation in anatomically complex regions of the head. The main drawbacks of Voxel-based techniques are that they demand substantial expert effort for the segmentation and are time-consuming.

2.2.3 Atlas-based

Atlas-based approaches rely on a set of pre-aligned MRI and CT image pairs, commonly referred to as atlases. In this framework, each MRI atlas is deformably registered to the target MRI of a new patient. The resulting deformation fields are then applied to the corresponding CT images within the atlas set, effectively mapping them into the target's anatomical space. The deformed CT images are subsequently combined or fused to produce a single synthetic CT (sCT) representation.[33] Early implementations of this method used a single average atlas, later advancements introduced multi-atlas frameworks, which significantly improved the resulting sCT quality by integrating information from multiple atlases. These systems often employed patch-based fusion and sparse-coding pattern recognition techniques to enhance local tissue correspondence and reduce registration errors. [34]

The primary limitation of atlas-based methods is their dependence on registration accuracy. It is computationally expensive and prone to topological errors when the target anatomy differs significantly from the atlas population. For example, in the pelvis, variable bladder and rectal filling can cause large, non-linear deformations that are difficult for standard approaches to resolve.[35]

2.3 Learning-based and deep learning

2.3.1 Supervised Learning

The advent of DL, such as Convolutional Neural Networks (CNNs) , revolutionized medical image synthesis by enabling the learning of hierarchical, non-linear mappings directly from data without the need for hand-crafted features. Han [36] was the first to introduce a deep CNN for synthetic CT generation from T1-weighted MRI. The proposed network architecture was adapted from models originally developed for object segmentation tasks. It operated on 2D MRI slices,

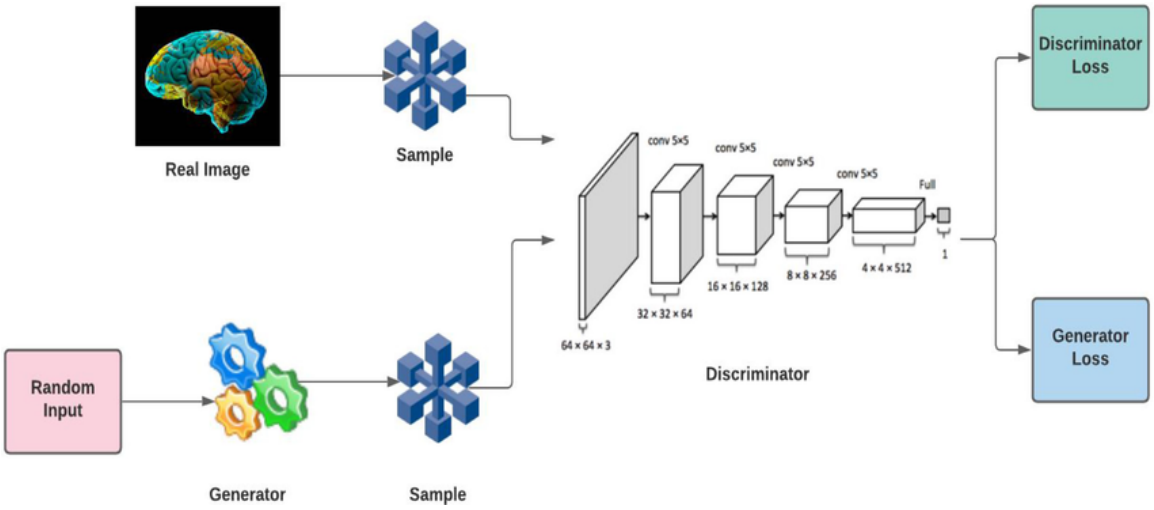


Figure 2.2: Generative Adversarial Network (GAN) concept [2]

converting each slice independently into its corresponding synthetic CT slice. However, this coarse processing often resulted in discontinuities and inconsistencies across adjacent slices in the reconstructed 3D sCT volume. Generative adversarial networks (GANs) were subsequently explored for MRI-to-sCT synthesis, beginning with the work of Nie et al. in 2017. [37] Their model processed small 3D patches (16 × 16 × 16 voxels) extracted from T1-weighted MRI volumes, producing synthetic CT with reduced blurring compared with fully convolutional network (FCN) baselines. Emami et al. [38] also demonstrated a GAN approach using 2D MRI slices and reported improved sCT quality relative to conventional CNN-based methods on the same dataset.Given the success of these learning-based MRI-to-CT synthesis methods, it is important to recognize that they rely on large paired MRI–CT datasets for training, which limits the ability to use additional MRI-only or CT-only scans from patients who lack both modalities.

Generative adversarial networks

GANs were first introduced by Goodfellow et al. [39], established a framework for estimating generative models via an adversarial process. The architecture consists of two simultaneous neural networks: a generator G that captures the data distribution, and a discriminator D that estimates the probability that a sample came from the training data rather than from G.The training process is formulated as a minimax two-player game. The generator takes a random noise vector z from

a prior distribution $p_z(z)$ and maps it to the output space. The discriminator aims to distinguish between real samples x drawn from the real data distribution $p_{data}(x)$ and fake samples generated by G. The standard objective function $V(D, G)$ is expressed as:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2}$$

In this standard formulation, minimizing the value function amounts to minimizing the Jensen-Shannon divergence between the real data distribution and the generated distribution. However, this approach often suffers from the vanishing gradient problem. When the discriminator is trained to optimality, it may stop providing sufficient gradient information for the generator to learn, leading to training instability. To address these stability issues, Mao et al. [40] proposed the Least Squares Generative Adversarial Network (LSGAN). Unlike the original GAN which utilizes a sigmoid cross-entropy loss (log loss), the LSGAN adopts a least squares loss function (L2 loss) for the discriminator10. Mathematically, minimizing this objective equates to minimising the Pearson χ^2 divergence11. The primary advantage of the LSGAN objective is that it penalizes fake samples based on their distance from the decision boundary. This forces the generator to produce samples that are closer to the decision boundary, effectively targeting fake samples that have a high margin. Consequently, LSGANs have been shown to generate higher quality images compared to standard GANs and successfully circumvent the vanishing gradient problem during the optimization process.

2.3.2 Unsupervised Learning

To make use of additional MRI-only or CT-only datasets from patients who were not scanned with both modalities, Wolterink et al. [41] employed a CycleGAN model to synthesize CT images from 2D brain MRI using unpaired training data. To further enhance structural consistency under unpaired conditions, Yang et al. [17] introduced a structure-constrained CycleGAN applied to the same type of data. While unsupervised frameworks like CycleGAN alleviate the need for paired data, they operate primarily on distribution matching rather than voxel-wise correspondence. Consequently, these models often suffer from geometric distortion and ”hallucinations”, where anatomical structures are invented or misplaced to satisfy the discriminator’s realism criteria rather than

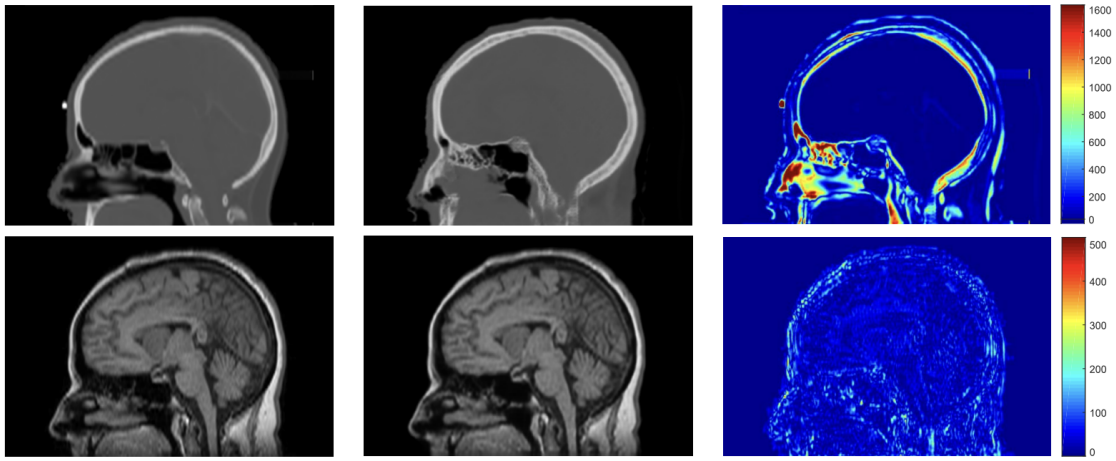


Figure 2.3: Visual example of a CycleGAN result. The left is a ground truth CT image (upper) and an input MR image (lower). The middle is a synthetic CT image (upper) and a reconstructed MR image (lower), and the right is the relative errors between the ground truth and synthetic CT images (upper) and the input and reconstructed MR images (lower). [3]

anatomical truth [42]. Furthermore, in radiotherapy planning, where pixel-level accuracy is critical for dose calculation, the lack of direct supervision can lead to unacceptable errors in Hounsfield Unit (HU) assignment, particularly at bone-air interfaces. In Figure 2.3, we can see the edge is poorly defined and exhibits geometric distortions, where the cortical bone boundary hallucinates non-existent structures.

2.3.3 Attention

Attention mechanisms in deep learning draw inspiration from the human visual system, which does not process an entire scene with equal intensity. Instead, human perception selectively concentrates on salient regions to extract relevant structural information while suppressing irrelevant background noise. Originally revolutionizing natural language processing (NLP) for machine translation tasks [43], attention mechanisms have since become a cornerstone in computer vision, enhancing performance in classification, segmentation, and image synthesis. Broadly, attention mechanisms can be categorized based on the dimension they recalibrate: channel attention focuses on ”what”

features are meaningful, such as distinguishing between texture and edge filters, while spatial attention focuses on "where" informative features are located. [44] While earlier approaches often relied on "hard" attention including cropping regions of interest, modern architectures utilize "soft" attention, which learns continuous weights between 0 and 1 to modulate feature maps in a differentiable manner. In the context of medical image analysis, specifically within U-Net architectures, the standard skip connections indiscriminately transfer low-level features responses—including background noise—to the decoder. To address this, Oktay et al. [45] proposed the Attention U-Net, which incorporates Additive Attention Gates (AGs). Unlike channel-based approaches like SE-Net [46] or CBAM [47] that recalibrate feature maps globally, the AG is a spatial mechanism designed to isolate relevant anatomical structures. The AG utilizes the coarser feature maps from the decoder path as a gating signal (g) to disambiguate the features coming from the encoder (x). By learning to align the gating signal with the encoder features, the network generates a spatial attention map that suppresses feature activations in irrelevant regions (such as background air) while highlighting the target organ boundaries. This allows the model to learn to focus on the target anatomy automatically without requiring explicit bounding box supervision or external cropping.

2.4 Recent learning-based advance

While GAN, CycleGAN and their variants have established a strong baseline for unsupervised medical image synthesis, the field has recently shifted toward addressing two primary limitations of these convolutional architectures: the lack of explicit structural constraints and the inability to model long-range dependencies. Recent innovations have focused on exploring non-adversarial generative paradigms and developing hybrid architectures.

2.4.1 Transformers and Diffusion Models

Beyond GANs, two deep learning paradigms have recently gained prominence in medical image analysis: Vision Transformers (ViTs) and Denoising Diffusion Probabilistic Models (DDPMs).

CNNs excel at extracting local features but often struggle to capture global context due to their limited receptive fields. Vision Transformers address this by using self-attention mechanisms to

model long-range dependencies across the entire image [48]. In synthesis tasks, this allows the model to understand the global spatial relationship between organs rather than just local texture statistics. A recent breakthrough in this domain is the work by Hu et al. [49], who recognized that standard CycleGAN generators often fail to preserve global structural integrity, they integrated a ViT-based module directly into the U-Net generator. In their architecture, the U-Net path extracts local anatomical features, while the embedded Transformer self-attention layers automatically prioritize information from distant spatial positions. This demonstrates the potential of Transformers to solve the ”local-only” bias of traditional GANs.

Diffusion models have emerged as a powerful alternative to GANs. By iteratively destroying data with noise and learning to reverse the process, diffusion models can potentially generate high-fidelity images with stable training dynamics, avoiding the mode collapse often seen in GANs.

To date, a major limitation of diffusion models has been their computational intensity, often restricting them to 2D slices. However, recent work has successfully extended this paradigm to volumetric data. Pan et al. [50] approach replaces the standard convolution backbone with a Shifted-Window Transformer V-Net (Swin-VNet) within the diffusion reverse process. This architecture allows the model to capture complex 3D anatomical relationships that 2D methods miss. This work strongly supports the viability of diffusion models as a robust, high-fidelity alternative to GANs for 3D medical image synthesis.

2.4.2 Hybrid and Structure-Guided Architectures

Recognizing that no single architecture solves all challenges, the current state-of-the-art has converged on hybrid approaches. These methods integrate the inference speed of GANs, the global context of Transformers, and the stability of Diffusion models, or incorporate explicit structural guidance to constrain the generation process. The MaskGAN framework [19], incorporates an auxiliary mask generator into the synthesis loop. Rather than relying solely on pixel-level adversarial loss, MaskGAN uses coarse anatomical masks to explicitly condition the generator, separating structural layout from texture generation.

These developments collectively suggest medical image synthesis is paving towards hybridization. Pure CNNs lack global context, pure Transformers require massive datasets, and pure Diffusion models are computationally expensive. Therefore, architectures that combine the efficiency of GANs (such as LSGAN) with the context-awareness of Attention mechanisms (such as Attention-UNet or ViT) represent the most promising direction for clinically viable, real-time synthetic CT generation.

Chapter 3

MAGnet: Mask-guided Generative Adversarial Network

3.1 Introduction

CT and MRI are jointly employed in radiotherapy planning because they contribute complementary strengths[51]: CT offers geometrically faithful electron-density information for accurate dose computation, whereas MRI provides rich anatomical and functional detail with superior soft-tissue contrast for delineating targets and organs at risk (OARs) [52]. These MRI advantages are especially salient in the pelvis, head and neck, and brain, where precise soft-tissue boundaries are essential for contouring [13]. At the same time, dual-modality workflows increase acquisition and registration burden and add ionizing radiation from CT[53]. Multimodal settings such as PET/CT, PET/MRI, and MR-guided radiotherapy[54] underscore the broader challenge of reconciling disparate contrast mechanisms; for example, PET/MRI lacks the CT data traditionally used for attenuation correction, and Dixon-based MR strategies struggle to represent bone accurately[55] because of large MR–CT contrast differences. These considerations have motivated interest in MRI-only pathways that recover CT-equivalent information directly from MRI.

Methods for MR-to-synthetic CT (sCT) generally fall into three categories: segmentation-based, atlas-based, and learning-based approaches[56]. Classical pipelines are vulnerable to registration drift and label errors [57], and segmentation-based "bulk-density" methods further require highly

reliable tissue classification. This tissue classification is inherently difficult on MRI because bone and nearby soft tissues often exhibit similar or sequence-dependent signal[58] levels across scanners and sites, leading to misassignment.

Deep Learning-based techniques have driven substantial progress. Early convolutional neural networks[59] (CNNs) [15] and generative adversarial networks[60] (GANs) produced convincing sCT in the head area [61], with documented gains for PET image quality [62]. More recent hybrid methods use adversarial diffusion models [63] to make the MRI-to-CT conversion look more accurate and realistic. Nonetheless, most end-to-end frameworks treat anatomy uniformly [64]. Such models frequently encounter trade-offs: cortical bone is prone to over-smoothing or bias, whereas soft-tissue texture may be lost or hallucinated. [65] These limitations are especially evident in pelvic imaging, where slender cortices, bowel gas, and motion further complicate prediction.

Against this backdrop, an effective sCT solution should encode anatomy-aware priors and respect the distinct statistics of bone and soft tissue [66], preserving bone geometry and CT numbers without compromising soft-tissue depiction. Building on this rationale, we pursue an MRI-only planning strategy that synthesizes CT from MRI while explicitly addressing these modality and anatomy-specific challenges across sites and sequences.

In this chapter, we propose MAGNet, a mask-guided, dual-branch GAN for synthesizing CT from MRI in anatomically complex regions. Our contributions are threefold: (i) a dual-branch, anatomy-conditioned GAN that allocates model capacity separately to bone structure and soft-tissue texture, mitigating the trade-offs common to uniform translators; (ii) a label-efficient guidance strategy that leverages TotalSegmentator[4, 20] masks to scale anatomy-aware conditioning without manual contouring; and (iii) comprehensive validation on public and internal datasets, demonstrating improved performance in challenging pelvic cases.

3.2 Method

This section provides a detailed description of the proposed MAGnet model, starting with an overview and subsequently examining each of its components.

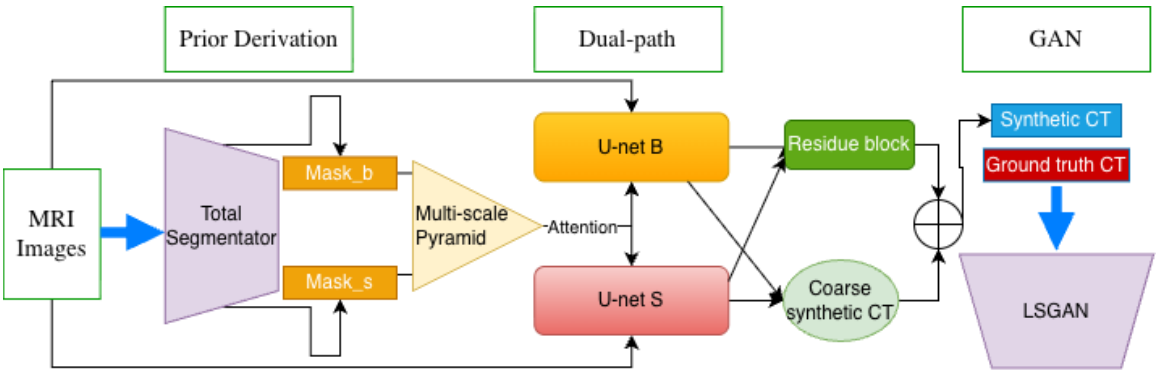


Figure 3.1: The network architecture of MAGnet

3.2.1 Model Overview

As illustrated in Fig. 3.1, the MAGNet framework operates in three stages: pyramidal prior derivation, anatomy-guided dual-path synthesis, and refinement-based compositing. Given an input MRI, bone and soft-tissue priors (M_b, M_s) are obtained via TotalSegmentator and constructed into a multi-scale pyramid to guide generation at varying resolutions [67]. Two specialized branches then predict s_{Bone} and s_{Tissue} using attention-gated mechanisms. Finally, the branch outputs are fused via a residual refinement module to correct boundary artifacts and form the final sCT.

3.2.2 Totalsegmentator

To derive high-fidelity anatomical priors without the prohibitive cost of manual delineation, this study utilizes TotalSegmentator, an open-source automated segmentation tool built upon the nnU-Net deep learning framework. Unlike traditional multi-atlas registration methods, which rely on computationally expensive deformable registration and often fail when patient anatomy deviates significantly from the atlas population, TotalSegmentator formulates segmentation as a direct voxel-wise classification task.

The model used was specifically the MRI-adapted variant, TotalSegmentator MRI, which extends the robust capabilities of the original CT-based model to T1- and T2-weighted magnetic resonance imaging sequences. The model was trained on a massive, heterogeneous dataset comprising over 1,100 scans, enabling it to robustly segment 80 distinct anatomical structures,including major organs, skeletal components, and vascular structures as shown in Figure 3.2 . This extensive

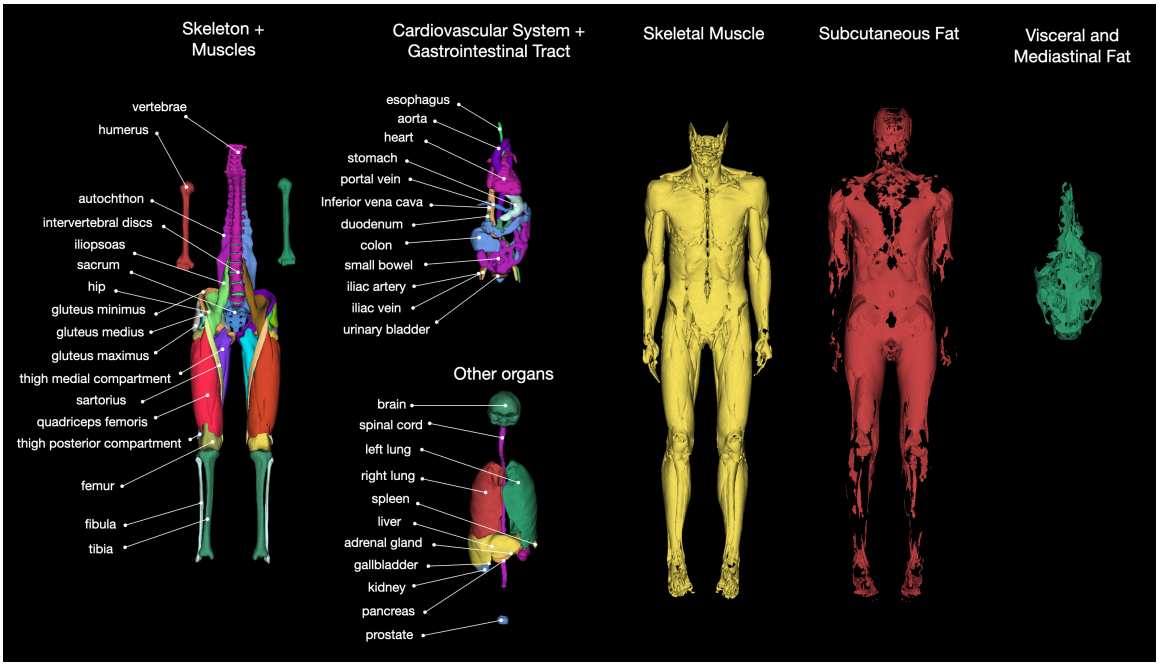


Figure 3.2: Overview of segmented anatomic structures from TotalSegmentator [4] .

training regime ensures that the segmentation remains reliable even in the presence of MRI-specific artifacts, such as bias field inhomogeneity and motion, which are common in pelvic imaging.

The selection of TotalSegmentator as the backbone for prior derivation is further motivated by its granular output space. Unlike standard tools that might only output a generic class for bone, this framework provides distinct labels for the femur, hip, sacrum, and individual vertebrae. This granularity allows for the precise aggregation of labels into custom semantic priors tailored to the synthesis task.

3.2.3 Derivation of Binary Priors

The raw output from the TotalSegmentator MRI model is a dense, multi-class label map $L \in \{0, \ldots, K\}^{H \times W \times D}$, where each voxel is assigned one of over 80 distinct anatomical class indices. While this granular segmentation provides rich semantic information, the proposed synthesis framework requires a simplified topological guide that specifically delineates the high-frequency structural components (cortical bone) from the heterogeneous soft-tissue regions. Therefore, we perform a semantic aggregation step to derive two binary occupancy priors: the bone mask (M_b) and the

soft-tissue mask (M_s).We define a subset of class indices $\mathcal{B} \subset \{1, \ldots, K\}$ corresponding to all osseous structures identified by the segmentation model. This subset includes the labels for the femur (left and right), hip joints, sacrum, and the complete lumbar and thoracic vertebral column present in the field of view. The binary bone prior M_b is constructed via an indicator function that assigns a value of 1 to any voxel v falling within this subset, and 0 otherwise. Simultaneously, we define the anatomical body contour Ω to exclude the background air and patient table, ensuring that the synthesis focuses solely on the relevant biological tissues. This region Ω is derived by aggregating all foreground labels predicted by TotalSegmentator. The soft-tissue prior M_s is then computed as the complement of the bone prior restricted to the body contour. Formally, the priors are derived as follows:

$$M_b(v) = \begin{cases} 1 & \text{if } L(v) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$M_s(v) = \begin{cases} 1 & \text{if } v \in \Omega \text{ and } L(v) \notin \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

This binary decomposition addresses the fundamental challenge in pelvic MRI-to-CT synthesis: the intensity ambiguity between cortical bone and air. In T2-weighted MRI, both tissues exhibit signal voids (hypointensity). By explicitly enforcing $M_b(v) = 1$ for bone voxels based on semantic context rather than signal intensity, the network is provided with a topological guarantee of where ossification should occur, regardless of the local signal dropout. The resulting masks serve as the inputs for the multi-scale pyramid construction described in the subsequent section.

3.2.4 Anatomy-Guided Dual-Branch Synthesis

To address the distinct generative challenges of cortical bone (high-frequency structure) versus soft tissue (heterogeneous texture), the MAGNet framework employs a Dual-Stream Generator architecture. As illustrated in Figure 3.1, two parallel U-Net generators ($\mathcal{G}_{bone}$ and $\mathcal{G}_{tissue}$) operate on the MRI input. While they share the same macro-architecture, they possess untied weights to learn specialized feature representations.To strictly localize the generative capacity of each branch,

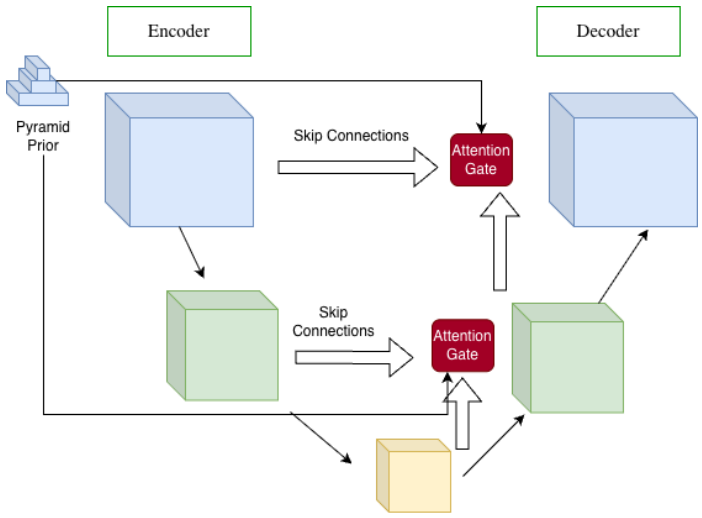


Figure 3.3: Detailed architecture of the anatomy-guided U-Net branch utilized in both the bone and soft-tissue streams. Unlike standard architectures, anatomical priors are not merely concatenated at the input but are structured into a Multi-Scale Pyramid. These downsampled masks $\{M^{(l)}\}$ are injected into the skip connections at every resolution level l to condition the Mask-Guided Attention Gates, ensuring that anatomical constraints are enforced consistently from coarse semantic features down to fine structural edges.

we introduce a Mask-Guided Attention mechanism within the skip connections.

Multi-Scale Mask Injection

In a standard U-Net, skip connections indiscriminately transfer low-level feature responses—including background noise and irrelevant anatomy—from the encoder to the decoder. To mitigate this, we utilize the Pyramidal Prior derived in 3.2.3. As shown in Figure 3.3, the downsampled masks $\{M^{(l)}\}$ are injected into the network at every resolution level l, providing spatial guidance from coarse semantic regions down to fine structural edges.

Attention Gate

We use the similar structure from Attention-Unet [45]. Unlike standard Additive Attention Gates which rely solely on feature correlations, our module introduces an explicit anatomical bias term as shown in Figure 3.4. Let x^l denote the input feature map from the encoder (the skip connection), and let g^l denote the gating signal from the decoder. To determine the attention coefficient, we align these features with the mask prior $M^{(l)}$ using learned linear transformations.

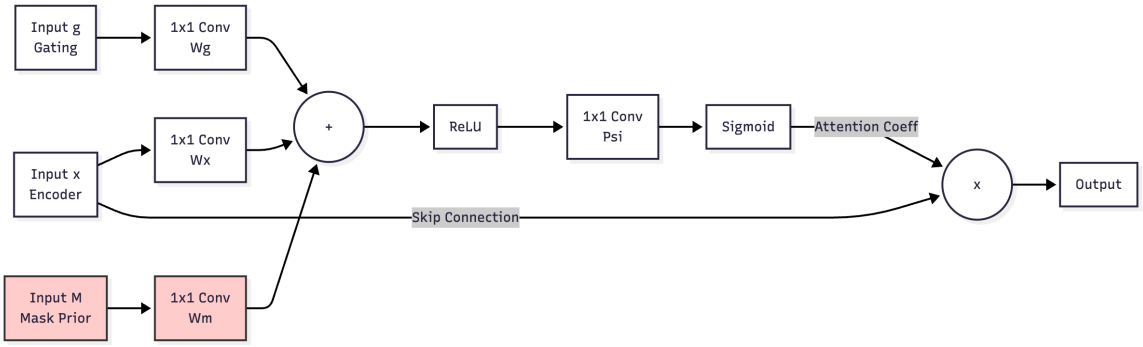


Figure 3.4: Schematic of the Attention Gate. The gating signal (g) from the decoder is fused with encoder features (x) and the projected mask prior (M). This anatomical injection biases the attention coefficient (α) toward the region of interest, spatially filtering the skip connection features to suppress irrelevant background noise.

The attention process proceeds as follows. First, the three inputs ($x^l, g^l, M^{(l)}$) are projected into a lower-dimensional intermediate space using 1×1 convolutions, denoted as W_x, W_g, and W_m, respectively. Second, the projected features are summed element-wise. This summation fuses the spatial "where" information from the mask with the semantic "what" information from the features. Third, the combined signal passes through a ReLU activation and a subsequent linear transformation Ψ (1×1 convolution) to collapse the channel dimensions. Lastly, a Sigmoid activation function (σ) normalizes the output to the range $[0, 1]$, producing the attention coefficient map α^l. Mathematically, this operation is defined as:

$$q_{att} = \Psi \left(\text{ReLU} \left(W_x x^l + W_g g^l + \mathbf{W_m M^{(l)}} \right) \right) \tag{5}$$

$$\alpha^l = \sigma_{sigmoid}(q_{att}) \tag{6}$$

Finally, the attention map filters the original encoder features via element-wise multiplication:

$$\hat{x}^l = x^l \cdot \alpha^l \tag{7}$$

By integrating the term $\mathbf{W_m M^{(l)}}$, the network learns to suppress feature activations in regions where the anatomical probability is zero (e.g., bone features appearing in the bladder), significantly reducing cross-region hallucination compared to standard concatenation methods.

Feature-Space Texture Consistency

While the proposed attention mechanism successfully localizes anatomical structures, soft-tissue regions in the pelvis exhibit high variance in texture and signal intensity compared to the relatively homogeneous cortical bone [68] . Standard pixel-wise losses such as L1 or MSE tend to smooth out these high-frequency textural details, leading to "waxy" or blurred soft-tissue predictions.To mitigate this, we incorporate a feature-space consistency constraint specifically for the soft-tissue branch $\mathcal{G}_{tissue}$. Let $\phi_j(\cdot)$ denote the feature activation map at the j-th layer of a pre-trained VGG-19 network. We enforce consistency between the generated soft tissue $\hat{S}$ and the ground truth soft tissue S_{gt} (where $S_{gt} = I_{CT} \odot M_s$) by minimizing the Euclidean distance in the feature space:

$$L_{tex} = \sum_{j \in \mathcal{J}} \frac{1}{C_j H_j W_j} ||\phi_j(\hat{S}) - \phi_j(S_{gt})||_2^2 \tag{8}$$

where $\mathcal{J}$ represents a set of intermediate ReLU layers that capture textural patterns at different scales. This term acts as a supervisor for the attention mechanism, penalizing any "hallucinated" bone textures that might bleed into the soft-tissue branch and ensuring that the visceral anatomy retains its characteristic HU distribution.

3.2.5 Refining composition

The final stage of the MAGNet framework addresses the integration of the disjoint branch outputs. Previous dual-path approaches typically rely on a naive linear summation such as $\hat{I} = \hat{B} + \hat{S}$ [68] to recombine the bone and tissue components. However, due to the discrete binary nature of the masks used for gating, this linear fusion often results in high-frequency seam artifacts or void pixels at the structural boundaries where M_b and M_s meet. To address this, we introduce a Residual Refinement Module, denoted as $\mathcal{R}$. As illustrated in the system overview (Figure 3.1), this module consists of a shallow stack of 3×3 convolutional layers with residual connections. Instead of attempting to regenerate the entire image, the module learns a sparse boundary correction map. The fusion process operates in two steps. First, a coarse prediction is generated via linear summation:

$$\hat{I}_{coarse} = \hat{B} \oplus \hat{S} \tag{9}$$

Next, the raw outputs of the bone and soft-tissue branches are concatenated channel-wise and passed through the refinement module to predict the residual error $\hat{I}_{res}$. The final synthetic CT ($\hat{I}_{sCT}$) is obtained by adding this correction to the coarse prediction:

$$\hat{I}res = \mathcal{R}(\text{Concat}(\hat{B}, \hat{S})) \tag{10}$$

$$\hat{I}_{sCT} = \hat{I}_{coarse} + \lambda res\hat{I}_{res} \tag{11}$$

By modeling the boundary error as a residual term, the network smooths the transitions between cortical bone and soft tissue without altering the accurate HU values within the homogeneous regions. This "Refinement via Residuals" strategy ensures that the final output is not merely a mosaic of two patches, but a cohesive, topologically continuous volume.

3.2.6 GAN Objective Loss Function

The optimization problem is formulated as a minimax game where the generator $\mathcal{G}$ (comprising the bone branch, soft-tissue branch, and refinement module) attempts to fool a discriminator network $\mathcal{D}$, while the discriminator attempts to distinguish between real CT images and synthesized sCTs.To ensure volumetric fidelity, structural realism, and texture consistency, we employ a composite objective function consisting of four terms: adversarial loss, voxel-wise fidelity loss, structural similarity loss, and texture consistency loss.

Least Squares Adversarial Loss

Standard GANs typically utilize a sigmoid cross-entropy loss function, which can lead to the vanishing gradient problem when the discriminator is confident. To stabilize training and generate higher quality images, we adopt the Least Squares GAN (LSGAN) objective [40] . The discriminator $\mathcal{D}$ minimizes the squared error between its predictions and the correct labels (1 for real, 0 for fake), while the generator minimizes the error of the discriminator on the synthesized images,trying to make them classified as 1.The discriminator loss $\mathcal{L}_{\mathcal{D}}$ is defined as:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}\mathbb{E}y \sim p_{data}(y)[(\mathcal{D}(y) - 1)^2] + \frac{1}{2}\mathbb{E}x \sim pdata(x)[(\mathcal{D}(\mathcal{G}(x)))^2] \tag{12}$$

The adversarial loss for the generator $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv} = \frac{1}{2}\mathbb{E}x \sim p_{data}(x)[(\mathcal{D}(\mathcal{G}(x)) - 1)^2] \tag{13}$$

where x represents the input MRI and y represents the ground truth CT (I_{CT}). This quadratic penalty enforces the generator to produce samples that lie close to the decision boundary, reducing mode collapse.

Voxel-Wise Fidelity

To ensure accurate HU assignment, which is critical for dose calculation in radiotherapy planning, we minimize the voxel-wise distance between the synthesized volume $\hat{I}_{sCT}$ and the ground truth I_{CT}. We utilize the L_1 norm rather than L_2, as L_1 encourages sharper boundaries and is less sensitive to outliers:

$$\mathcal{L}_{L1} = \mathbb{E}x, y[||y - \hat{I}_{sCT}||_1] \tag{14}$$

Structural Similarity Loss

While L_1 loss ensures intensity accuracy, it treats pixels independently and can sometimes result in blurry images. To preserve local structural information and contrast, we incorporate the Structural Similarity Index (SSIM) as part of the loss function. The SSIM loss maximizes the perceptual similarity between the local window patches of the synthesized and real images:

$$\mathcal{L}_{SSIM} = 1 - \text{SSIM}(I_{CT}, \hat{I}_{sCT}) \tag{15}$$

Total Generator Objective

The final objective function for the generator is a weighted sum of the adversarial loss, the reconstruction losses, and the texture consistency loss:

$$\mathcal{L}_G = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{SSIM}\mathcal{L}_{SSIM} + \lambda_{tex}\mathcal{L}_{tex} \tag{16}$$

where $\lambda_{adv}, \lambda_{L1}, \lambda_{SSIM}$, and λ_{tex} are hyperparameters controlling the relative importance of each term. This multi-term objective ensures that the synthesized CT images are not only visually realistic but also geometrically accurate and texturally consistent with the reference anatomy.

3.2.7 Optimization Strategy

The training proceeds in an alternating iterative schedule. Let θ_G and θ_D denote the parameters of the generator and discriminator, respectively. In each training iteration, we sample a batch of paired images (I_{MR}, I_{CT}) and perform the following two-step update: (i) We first compute the synthetic output $\hat{I}_{sCT} = \mathcal{G}(I_{MR})$. The discriminator parameters θ_D are updated to minimize $\mathcal{L}_{\mathcal{D}}$ (Eq 12) via gradient descent, while keeping the generator parameters θ_G fixed. This step optimizes the discriminator's ability to distinguish between the ground truth and the current synthetic estimate. (ii) In the subsequent step, the generator parameters θ_G are updated to minimize the total objective $\mathcal{L}_G$ (Eq 16), while keeping θ_D fixed. By minimizing the adversarial component $\mathcal{L}_{adv}$, the generator learns to produce images that the frozen discriminator classifies as real, simultaneously optimizing for voxel-wise and structural fidelity via $\mathcal{L}_{L1}$ and $\mathcal{L}_{SSIM}$. We utilize the same data batch for both updates to minimize gradient variance and ensure stable convergence.

Chapter 4

Experiment and Results

4.1 Overview

In this work, we investigate the efficacy of the proposed MAGNet framework for synthesizing pseudo-CT images from MRI data, specifically targeting the anatomically complex pelvic region. To validate the hypothesis that anatomy-guided synthesis yields superior geometric and textural fidelity compared to uniform translation methods, the following set of experiments was performed:

- We evaluate MAGNet against a comprehensive suite of state-of-the-art baselines, representing three distinct architectural paradigms:
    - Paired (Pix2Pix) and unpaired (CycleGAN) convolutional frameworks.
    - The transformer architecture (ResViT), to assess the comparative benefit of vision transformers in capturing long-range dependencies versus the proposed local attention mechanism.
    - MaskGAN, to directly compare the efficacy of the proposed multi-scale attention injection against alternative mask-guidance strategies.
- Conducting ablation studies by isolating the contributions of the key architectural components including the Pyramidal Prior, the Mask-Guided Attention, and the Residual Module to quantify their individual impact on model performance.

- Validating the robustness of the trained networks across two distinct cohorts: a standardized public benchmark (Gold Atlas) and a heterogeneous internal clinical dataset (LMU Munich).

The robustness of the trained networks was analyzed using clinical data prepared specifically for this study. This chapter describes the implementation details of the experimental pipeline, from the dataset curation and preprocessing strategies to the specific training hyperparameters. The quality of the sCT generation was assessed using voxel-wise metrics, including Normalized Mean Square Error (NMSE), as well as perceptual metrics such as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). To provide a qualitative assessment of geometric fidelity, we also present difference maps that visualize the spatial distribution of reconstruction errors.

4.2 Dataset

Gold Atlas Dataset

The primary benchmark employed for validation in this study is the Gold Atlas dataset [69] , a multi-institutional collection of male pelvic imaging data specifically curated to facilitate the development of sCT and automated segmentation algorithms. The cohort consists of 19 male patients recruited from three different radiotherapy departments in Sweden, all of whom were referred for curative radiotherapy for prostate or rectal cancer. To ensure the dataset represented typical treatment scenarios while avoiding extreme anatomical distortions, patients with locally advanced tumors (prostate cT3-4 or rectal cT4) were excluded from the study. For every subject, a planning CT and paired Magnetic Resonance Imaging (MRI) sequences (T1-weighted and T2-weighted) are available. Furthermore, the dataset includes expert consensus delineations for nine pelvic organs, derived from five independent observers, as well as deformably registered CT images to facilitate voxel-wise comparison.

This dataset was selected for validation due to its high clinical realism and multi-site variability. All patients were scanned using standard clinical fixation devices on flat table tops, ensuring that the body contour and organ positioning reflect actual radiotherapy treatment conditions rather than diagnostic imaging setups. To prevent the model from overfitting to the noise characteristics of a single device, the data was acquired across three different institutions using scanners from different

manufacturers. Image acquisition protocols varied by site to capture inter-scanner variability. All CT scans were acquired as part of the clinical routine for treatment planning. At Site 1, images were acquired on a Siemens Somatom Definition AS+ with a slice thickness of 3 mm; Site 2 utilized a Toshiba Aquilion with a slice thickness of 2 mm; and Site 3 employed a Siemens Emotion 6 with a slice thickness of 2.5 mm.

Corresponding MRI scans were performed with the patients in the treatment position. For this study, we utilized the T2-weighted sequences, which provide superior soft-tissue contrast for pelvic anatomy. These were acquired using diverse hardware configurations: Site 1 used a GE Discovery 750w with a Fast Recovery Fast Spin-Echo (FRFSE) sequence and 3 mm slice thickness; Site 2 used a Siemens Skyra with a Turbo Spin-Echo (TSE) sequence and 2 mm slice thickness; and Site 3 used a GE Signa PET/MR with an FRFSE sequence and 2.5 mm slice thickness. This heterogeneity provides a robust test bed for assessing the generalization capabilities of the proposed synthesis framework.

LMU Munich

To evaluate the model's performance in a heterogeneous clinical environment, thanks to the Prof. Dr. Susanne Mayer's Lab, we have been granted access to a large scale internal dataset in collaboration with the Musculoskeletal University Center at Ludwig Maximilian University (LMU) of Munich. Unlike the standardized Gold Atlas benchmark, this cohort represents a broad clinical distribution, featuring significant variations in patient anatomy, pathology, and scanner vendors. The initial database query yielded 1,219 patients who underwent pelvic imaging around 2015. From this pool, we identified 370 cases with paired axial T2-weighted MRI and CT scans acquired within a four-week interval. To ensure high-quality training data for the synthesis task, we applied strict exclusion criteria. Patients with metallic hip implants, which cause severe streak artifacts in CT and signal voids in MRI, were removed, along with cases exhibiting significant motion artifacts or incomplete fields of view. The final cohort consists of 170 paired volumes. To prevent data leakage, the dataset was split at the patient level into 120 training, 20 validation, and 30 testing subjects. All data were fully anonymized in accordance with the General Data Protection Regulation (GDPR), and the study was approved by the institutional ethics committee.

The imaging data reflects the diversity of a high volume clinical center, utilizing a wide range of hardware generations and acquisition protocols. CT scans were acquired using a diverse fleet of scanners from three major manufacturers, ensuring the model is robust to variations in reconstruction kernels and noise profiles. Based on the DICOM metadata, the distribution includes systems from Siemens Healthineers (Somatom Definition AS+, Somatom Force, Somatom Edge, Sensation 64), GE Medical Systems (LightSpeed VCT, Discovery CT750 HD), and Philips Medical Systems (Brilliance 64). The acquisition protocols varied by clinical indication, with tube potentials ranging from 100 to 140 kVp (median: 120 kVp). Slice thickness varied significantly across the cohort, ranging from high-resolution 0.6 mm reconstructions to standard 5.0 mm slices, with a median slice thickness of 2.0 mm. Reconstruction kernels included both standard soft-tissue filters (e.g., Siemens B30f, GE Standard) and sharper bone-enhancing kernels (e.g., Siemens B60f), providing a challenging domain adaptation task for the synthesis network.

Similarly, the MRI scans were performed on both 1.5 T and 3.0 T systems, primarily utilizing Siemens Magnetom Aera, Skyra, and Vida scanners. For this study, we utilized the T2-weighted Turbo Spin Echo (TSE) sequences without fat suppression, as they provide the optimal contrast between the hyperintense bladder/fat and the hypointense cortical bone. Images were typically acquired with an in-plane resolution of 0.6 to 0.8 mm and a slice thickness ranging from 3.0 mm to 4.0 mm, reflecting standard clinical protocols for pelvic examinations.

4.3 Data preprocessing

To ensure data consistency across multi-site cohorts and to mitigate the fundamental domain shift between MRI signal intensities and CT Hounsfield Units , a standardized preprocessing pipeline was applied to all paired volumes prior to network training.

Geometric Standardization and Artifact Removal

Since the raw data originated from multiple scanners with varying fields of view and pixel spacings, geometric normalization was a prerequisite. All MRI and CT volumes were resampled to a common isotropic resolution of $1.0 \times 1.0 \times 1.0$ mm using cubic spline interpolation. This step

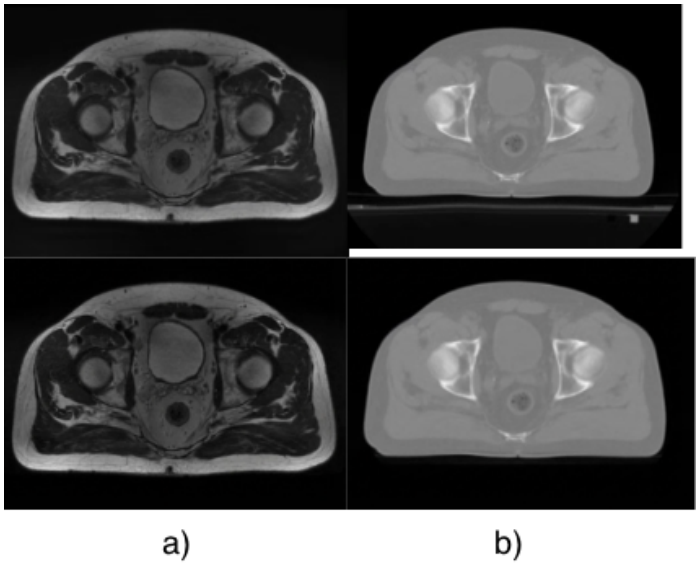


Figure 4.1: Visualization of the data preprocessing pipeline. **Top Row**: Representative raw MRI (a) and CT (b) axial slices prior to processing. Note the presence of non-anatomical artifacts, including the patient support table (couch) and background noise. **Bottom Row**: The corresponding volumes following geometric standardization and automated ROI extraction. The application of the body mask has successfully suppressed the patient table and background air, isolating the anatomical region of interest for network training.

guarantees that the convolution filters in the network operate on physical features of consistent spatial scale, facilitating robust feature extraction across patients with varying body sizes.Following resampling, we addressed the presence of non-anatomical artifacts common in clinical scans, such as the patient support table, fixation devices, and substantial background air. These features introduce high-contrast edges that are irrelevant to the anatomical synthesis task and can destabilize adversarial training. To mitigate this, an automated Region of Interest (ROI) extraction strategy was employed. We utilized the external body contour Ω, derived via the TotalSegmentator framework, to generate a binary body mask. This mask was applied element-wise to both the MRI and CT volumes. All voxels lying outside the biological envelope were strictly set to a constant padding value, effectively suppressing the patient table and background noise to force the network to focus exclusively on anatomical structures.

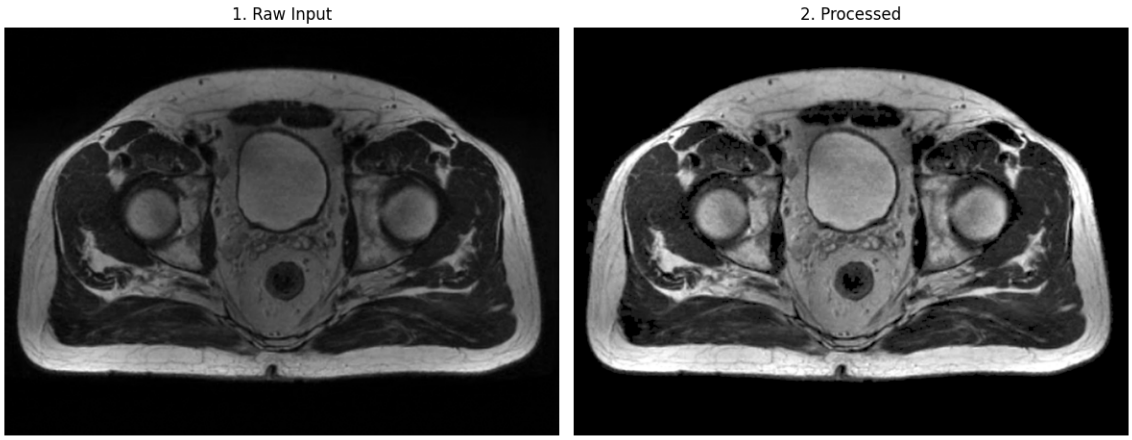


Figure 4.2: Impact of N4 bias field correction on T2-weighted MRI data. Left: Raw input volume exhibiting characteristic low-frequency intensity inhomogeneity (shading artifacts), where signal intensity varies spatially across the field of view due to magnetic field imperfections. Right: The corrected volume following the application of the N4ITK algorithm.

Modality Specific Intensity Normalization

Given that MRI and CT data exist in fundamentally different intensity domains, distinct normalization strategies were required for each modality.For the T2-weighted MRI volumes, which suffer from low-frequency intensity inhomogeneity caused by magnetic field variations, we first applied the N4ITK bias field correction algorithm [70]. Following this correction, a robust percentile-based normalization was applied to standardize the arbitrary signal intensities across different scanner vendors. Specifically, the signal intensities were divided by the 99^{th} percentile of the foreground body voxels. This approach scales the prominent soft tissues to the $[0, 1]$ range while preserving the relative contrast of hyperintense structures—such as fluid in the bladder—without the information loss associated with hard clipping.Conversely, the CT data possesses a quantitative physical meaning based on electron density. Therefore, preprocessing focused on windowing the dynamic range to the relevant radiological densities. The raw Hounsfield Units were clipped to the interval of $[-1000, 1500]$ HU. This window encompasses the full spectrum of tissues relevant for radiotherapy dose calculation, from air pockets to dense cortical bone, while removing high-intensity metallic artifacts. Finally, these clipped values were linearly scaled to the range $[-1, 1]$ to align with the output space of the generator's tanh activation function.

4.3.1 Evaluation Metrics

To provide a comprehensive assessment of the synthesized CT images, we employed a diverse set of quantitative metrics evaluating both voxel-wise intensity fidelity and perceptual structural similarity. Let I_{CT} denote the ground truth CT volume and $\hat{I}_{sCT}$ denote the synthetic pseudo-CT generated by the model. Both volumes are normalized to the HU scale for calculation.

Voxel-Wise Fidelity

To quantify the precision of the synthesized tissue densities, we utilized the Mean Absolute Error (MAE). While the Mean Squared Error (MSE) is a common metric in general image restoration, it penalizes outliers quadratically, often allowing small structural misalignments at organ boundaries to disproportionately dominate the error metric. In the context of medical imaging, where edge preservation is critical, the MAE utilizes the L1-norm to provide a linear quantification of error that is robust to outliers and more interpretable. The MAE is defined as the average absolute difference between the ground truth and synthesized voxels:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{I}sCT(i) - ICT(i)\right| \tag{17}$$

where N represents the total number of voxels within the evaluation mask. In CT, the MAE serves as a direct proxy for the accuracy of the synthesized electron density. An MAE value calculated on a normalized tensor between 0 and 1 can be linearly mapped back to the HU scale to estimate physical error. For instance, given a typical dynamic range spanning 4000 HU (from -1000 HU to +3000 HU), a normalized MAE allows us to assess whether the synthesis error falls within the acceptable variance for soft tissue differentiation or bone density calculation.

Perceptual and Structural Quality

To evaluate the visual realism and structural preservation achieved by the generator, we employed two widely accepted perceptual metrics:Peak Signal-to-Noise Ratio (PSNR):

- PSNR measures the ratio between the maximum possible signal power (MAX_I^2) and the power

of the reconstruction error (MSE), expressed in decibels (dB). A higher PSNR indicates superior signal fidelity.

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}_I^2}{\text{MSE}} \right) \tag{18}$$

- Structural Similarity Index (SSIM): SSIM is a full-reference metric that assesses the perceived structural quality by comparing local windows based on their luminance, contrast, and structure. SSIM is crucial for medical synthesis as it specifically penalizes artifacts like blurring that affect structural integrity. The final SSIM value ranges from 0 (no similarity) to 1 (perfect similarity).

4.4 Baselines and Comparison models

Standard Conditional GANs

These models represent the foundational approaches in medical image-to-image translation. We utilized the official implementations adapted for our paired training data.

- Pix2Pix [21] . This serves as the primary benchmark for paired image translation. We employ its standard architecture on which involved a U-Net Generator with skip connections and a PatchGAN Discriminator. Pix2Pix is tested to establish the maximum fidelity achievable solely through voxel-wise pairing without anatomical guidance.

- CycleGAN [71] . This model tests unpaired domain translation, though we adapt it for paired data for a direct performance comparison. We use the Pytorch implementation of ResNet Generator (9 blocks) and a PatchGAN Discriminator, maintaining the standard configuration for its loss terms (λ_{cyc}).

State-of-the-Art Architectures

These models test the performance of MAGNet against recent innovations in structural encoding and complex attention mechanisms.

- ResViT [72]. This hybrid architecture integrates a Vision Transformer (ViT) module within a convolution network to capture long-range dependencies and global context. We include ResViT to test whether the proposed local Attention Gate mechanism can outperform global self-attention strategies in the context of pelvic synthesis.

  MaskGAN [19]. This model directly competes with MAGNet's philosophical foundation by utilizing an alternative mask-guided synthesis strategy. We include MaskGAN to directly compare the efficacy of our multi-Scale attention injection against their mask-conditioning methods.

4.4.1 Training Configuration

The proposed MAGNet framework and all comparative baseline models were trained using the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.999$). To ensure a fair comparison, all models utilized the same training schedule of 150 epochs. The experiments were conducted on computer cluster consisting of NVIDIA V100 and A100 GPUs. Table 4.1 summarizes the key hyperparameters and the weight distribution for the composite loss functions.

Parameter	**MAGNet**	**Pix2Pix**	**CycleGAN**	**ResViT**	**MaskGAN**
Base Learning Rate (LR)	1×10^{-4}	2×10^{-4}	2×10^{-4}	1×10^{-4}	1×10^{-4}
Batch Size	16	16	16	8	16
λ_{Adv} (LSGAN)	1	1	1	N/A	1
$\lambda_{\mathbf{L1}}$ (Voxel Fidelity)	100	100	N/A	100	100
$\lambda_{\mathbf{SSIM}}$ (Structure)	10	10	10	10	10
$\lambda_{\mathbf{Cyc}}$ (Cycle Loss)	N/A	N/A	10	N/A	N/A
$\lambda_{\mathbf{Tex}}$ (Feature-space Loss)	10	N/A	N/A	N/A	10
$\lambda_{\mathbf{Res}}$ (Refinement Module)	1	N/A	N/A	N/A	N/A

Table 4.1: Key Hyperparameter and Loss Weight Configuration for MAGNet and Baseline Models.

4.5 Result and Discussion

To validate the efficacy of the proposed MAGNet framework, we performed a comprehensive evaluation on both the public Gold Atlas benchmark and the internal LMU clinical cohort. The

Model	MAE (HU) ↓	PSNR (dB) ↑	SSIM ↑
CycleGAN[71]	55.8 ± 4.9	17.3 ± 1.5	0.644 ± 0.038
Pix2Pix[21]	53.2 ± 4.2	22.1 ± 1.7	0.842 ± 0.026
ResViT[72]	41.5 ± 3.8	24.2 ± 1.3	0.891 ± 0.012
MaskGAN[19]	38.9 ± 4.1	23.8 ± 1.5	0.905 ± 0.021
MAGNet (Ours)	$\mathbf{35.4 \pm 3.9}$	$\mathbf{25.2 \pm 1.2}$	$\mathbf{0.922 \pm 0.016}$

Table 4.2: Quantitative Comparison on Gold Atlas (Public) Dataset. Values are reported as Mean ± Standard Deviation. Best results are bolded.

results are organized as follows: first, we present the quantitative comparison against state-of-the-art baselines; second, we analyze the contribution of individual architectural components through an ablation study; and finally, we provide a qualitative visual analysis of the synthesized anatomies.

4.5.1 Quantitative Comparative Analysis

Performance on Gold Atlas Benchmark

Table 4.2 summarizes the quantitative performance of MAGNet and the baseline models on the Gold Atlas testing set. The proposed framework achieved the lowest MAE of 35.4 HU and the highest PSNR of 25.2 dB.

The results highlight a clear hierarchy in synthesis quality. The unpaired CycleGAN baseline performed poorly, yielding a PSNR of only 17.3dB. This low score, combined with the highest standard deviation in MAE (± 4.9 HU), indicates that the model frequently failed to converge on a geometrically accurate solution. The paired Pix2Pix model improved substantially upon this baseline (PSNR 22.1 dB), suggesting the importance of supervised pixel-level correspondence.

A performance increase was observed with the introduction of advanced architectures. The transformer-based ResViT achieved a PSNR of 24.2 dB, benefiting from its global attention mechanism. Interestingly, while MaskGAN achieved a high Structural SSIM of 0.905 confirming that explicit anatomical conditioning is valuable, it yielded a slightly lower PSNR of 23.8 dB than ResViT. This suggests that while MaskGAN effectively captures boundaries, its reliance on unpaired adversarial training and mask-guided constraints may struggle to regress precise HU within complex texture regions compared to the transformer-based approach.

MAGNet outperformed all baselines across all metrics, achieving a SSIM of 0.922 and a PSNR

Model	MAE (HU) ↓	PSNR (dB) ↑	SSIM ↑
CycleGAN	61.5 ± 5.2	16.7 ± 1.9	0.630 ± 0.041
Pix2Pix	57.4 ± 4.5	21.4 ± 1.6	0.827 ± 0.028
MAGNet (Ours)	$\mathbf{38.5 \pm 4.1}$	$\mathbf{23.9 \pm 1.3}$	$\mathbf{0.862 \pm 0.022}$

Table 4.3: Quantitative Comparison on LMU Munich (Internal) Dataset. Values are reported as Mean $\pm$ Standard Deviation. Best results are bolded.

of 25.2 dB. By leveraging priors from TotalSegmentator and combining them with the pixel-level precision of paired supervision, MAGNet effectively bridges the gap between these competing paradigms. It retains the structural rigidity of MaskGAN while surpassing the intensity accuracy of ResViT, demonstrating that the proposed attention gates effectively resolve the structural ambiguities that limit current state-of-the-art competitors.

Performance on Internal LMU Clinical Cohort

Following the evaluation on the standardized Gold Atlas benchmark, we assessed the model's robustness on the internal LMU Munich cohort. As detailed in Section 4.2, this dataset represents a significantly more challenging domain due to its heterogeneity, featuring data acquired from multiple scanner vendors with varying acquisition protocols and reconstruction kernels. This evaluation aims to determine if the anatomy-guided priors can maintain synthesis fidelity in a realistic, multi-center clinical environment where standard distribution-matching methods often fail.

Table 4.3 presents the quantitative results for the internal cohort. It is important to note that ResViT and MaskGAN, which were evaluated on the public dataset, could not be deployed on this internal cohort. Strict data governance policies associated with the clinical collaboration restricted the data retention period, rendering the dataset inaccessible for retrospective analysis during the phase when these advanced architectures were integrated later into the study. Consequently, comparisons on the internal dataset are focused on the foundational paired (Pix2Pix) and unpaired (CycleGAN) baselines.

The results on the internal cohort mirror the hierarchical trends observed in the public benchmark. CycleGAN, lacking paired supervision, struggled to converge on the heterogeneous data, yielding the lowest fidelity with a PSNR of 16.7 ± 1.9 dB and SSIM of 0.63. Pix2Pix provided a stronger baseline with a PSNR of 21.4 ± 1.6 dB. MAGNet demonstrated superior generalization

Feature	Gold Atlas (Public)	Internal LMU Dataset
Number of Patients	19	1219 total studies
Used Subset	All	370 axial T2-weighted studies
Region	Male Pelvis	Abdomen & Pelvis
Registration	Deformable; Verified	Standard Clinical
MR Orientations	Axial only	Axial, Coronal, Sagittal
MR Sequences	T2-weighted (2.5mm)	T1, T2, T1-FS (Axial T2: 3mm)
CT Slice Thickness	2.0 – 3.0mm	1.5 – 3.0mm
Paired Data	Yes	Yes

Table 4.4: Comparison of Dataset Characteristics.

capabilities, achieving a PSNR of 23.9 ± 1.3 dB and an SSIM of 0.862. This margin is particularly revealing given the heterogeneity of the internal cohort. Conventional translation models such as Pix2Pix often struggle to learn a single mapping that generalizes across the diverse soft-tissue contrasts produced by different scanner vendors, resulting in degraded performance [73]. In contrast, MAGNet's superior fidelity indicates that incorporating explicit anatomical priors provides a critical geometric anchor for the generation process. By constraining the solution space through anatomical structure rather than relying solely on intensity correlations, the model preserves structural consistency even when input signals vary across scanners.

4.5.2 Analysis of Generalization Gap

A comparison between the Gold Atlas and LMU dataset results reveals a moderate but expected performance gap. On the standardized Gold Atlas dataset, MAGNet achieved a peak fidelity of 25.2 dB. On the heterogeneous internal cohort, performance settled at 23.9 dB.

This difference highlights the distinction between idealized benchmarks and real-world deployment, as detailed in Table 4.4. The Gold Atlas dataset consists of a small, curated set of 19 male pelvis patients, where ground truth CTs were deformably registered and verified by five independent radiologists to minimize geometric mismatch. In contrast, the internal evaluation utilized a cohort of 170 patients (selected from a larger screening of 370 Axial T2 studies). While matched in sequence type of Axial T2, this internal cohort introduces significant scanner variability absent

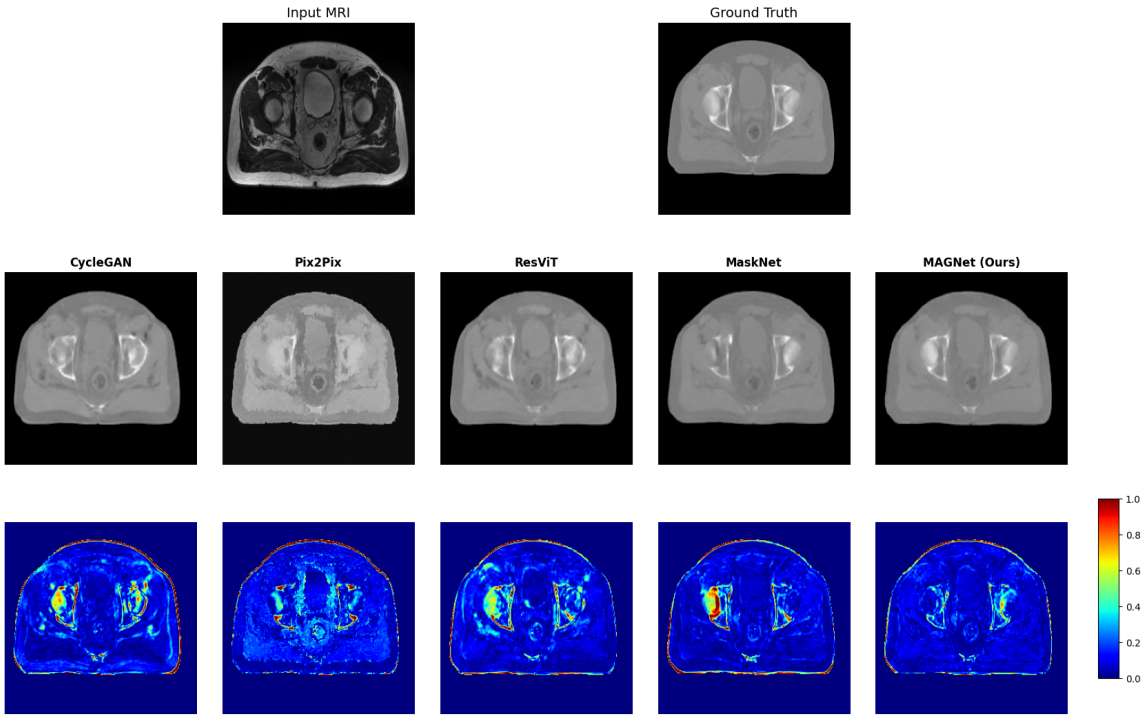


Figure 4.3: Qualitative comparison of synthetic CT generation methods on a representative axial slice from the Gold Atlas dataset. **Top Row:** The input MRI and the Ground Truth CT. **Middle Row:** Synthetic CT predictions generated by the baseline models (CycleGAN, Pix2Pix), advanced architectures (ResViT, MaskNet), and the proposed MAGNet framework. **Bottom Row:**Absolute error maps where dark blue indicates low error and red indicates high error.

in the public benchmark, utilizing a diverse fleet of Siemens, GE, and Philips systems with varying reconstruction kernels.

Furthermore, unlike the verified registration of the benchmark, the internal cohort relies on standard clinical alignment where natural non-rigid anatomical variations remain. Despite this domain shift and the lack of artificial deformable registration, MAGNet maintained high structural similarity, demonstrating that the anatomical priors effectively stabilized the generation process even when pixel-wise intensity mapping became more challenging.

4.5.3 Qualitative Visual Analysis

To complement the quantitative metrics reported in the previous sections, we conducted a comprehensive visual inspection of the sCT volumes to assess anatomical realism, texture fidelity, and the spatial distribution of reconstruction errors. Figure 4.3 presents a representative axial slice from

the Gold Atlas dataset, displaying the input MRI, the Ground Truth CT, and the sCT predictions generated by the comparative methods, alongside their corresponding absolute error maps.

From the second row, a clear hierarchy in synthesis quality was revealed which supports the quantitative findings. The foundational baseline methods exhibit characteristic artifacts associated with unconstrained generation. CycleGAN suffers from significant geometric distortions and blurring, particularly around the femoral heads and the acetabulum, failing to preserve the sharp high-frequency edges of the cortical bone required for accurate dose calculation. Furthermore, the model exhibits a clear tendency to hallucinate non-existent structural boundaries as seen within the femoral head on the right side of the image. CycleGAN generates erroneous internal edges and high-intensity artifacts that lack any anatomical basis in the ground truth.This failure mode provides visual confirmation of the theoretical limitations inherent to CycleGAN's unsupervised nature. As discussed in Section 2.3.2, it relies on cycle-consistency loss rather than direct paired supervision to preserve structural information. Without a voxel-wise objective to anchor the synthesis to the ground truth anatomy, the generator is free to invent structures that satisfy the discriminator's textural requirements but violate geometric truth. Consequently, the model prioritizes perceptual distribution matching over structural fidelity, leading to plausible on paper but anatomically incorrect hallucinations in complex regions like trabecular bone.

Pix2Pix, while achieving better global structural alignment due to paired supervision, introduces high-frequency artifacts within the homogeneous soft-tissue regions such as the gluteal muscles and bladder. This results in a texture that deviates markedly from the smooth, uniform HU distribution characteristic of real CT scans.

The advanced architectures, ResViT and MaskNet, demonstrate a substantial improvement in image quality over the foundational baselines. ResViT achieves a significantly more consistent and homogeneous soft-tissue texture compared to Pix2Pix, likely a result of the global context modeling provided by the vision transformer which enforces spatial coherence across the volume. However, it exhibits a tendency toward over-smoothing, particularly at fine bony edges where pixel-perfect precision is critical. MaskNet produces sharp cortical boundaries comparable to the ground truth, validating the efficacy of segmentation-guided synthesis. However, a critical limitation is observed in the internal osseous structures. The interior trabecular bone of the left femoral head appears

significantly under-dense. This suggests that while MaskNet correctly localizes the cortical shell via the mask, the generator fails to accurately regress the internal density of the spongy bone, leaving it visually hollow.

In contrast, the proposed MAGNet framework yields the most anatomically faithful reconstruction among the evaluated cohorts. By combining the local feature refinement of the dual-path architecture with explicit mask guidance, it retains the sharp bone definition seen in MaskNet while correcting the density drop-off observed in the femoral interior. The dual-branch design allows the model to allocate distinct generative capacities to different tissue types, resulting in preserved trabecular details inside the femoral heads without the hollowing artifacts that affect the single-stream MaskNet architecture.

The error maps presented in the third row provide a spatial visualization of the reconstruction error, where blue indicates low error and red indicates high error, further elucidating the failure modes of each architecture. The error map for Pix2Pix exhibits a high-frequency spatial variance distribution scattered throughout the soft-tissue regions, confirming that the standard convolutional generator struggles to regress stable HU for large, homogeneous tissue classes. CycleGAN displays broad, contiguous regions of high error around structural interfaces, indicating fundamental geometric misalignment. While ResViT improves textural consistency, its error map is dominated by distinct error outlines tracing the contours of the pelvic bones. This suggests that while the Vision Transformer successfully models global context, it struggles with pixel-perfect alignment at high-contrast interfaces, leading to edge specific residuals.

Most critically, the difference maps highlight the superior internal bone fidelity of MAGNet compared to the competing mask-guided approach. The MaskNet error map reveals localized hot spots of high error within the internal regions of the femoral heads, correlating directly with the missing density observed in the synthesized image. This indicates that the model under-predicted the HU values of the trabecular bone, effectively treating it as soft tissue. In contrast, MAGNet demonstrates a significantly cooler error profile within these osseous structures. This reduction in internal bone error validates the contribution of the specialized bone synthesis branch, which effectively learns the distinct textural statistics of trabecular bone. The resulting error map for

MAGNet is predominantly dark blue, indicating near-zero error across the majority of the soft-tissue volume, with residual errors strictly confined to the immediate periphery of the cortical bone.

4.5.4 Ablation Studies

To systematically validate the effectiveness of the proposed framework, we conducted ablation studies on the Gold Atlas dataset. Following the experimental methodology established in other literature [68, 72] , we isolated the contributions of the specific objective functions and the architectural modules to quantify their individual impact on synthesis fidelity.

Effectiveness of loss function

Loss Combination	**PSNR (dB)** ↑	**SSIM** ↑
$L_{adv} + L_{L1}$	24.5 ± 1.1	0.908 ± 0.014
$L_{adv} + L_{L1} + L_{SSIM}$	24.9 ± 1.3	0.916 ± 0.021
$L_{adv} + L_{L1} + L_{SSIM} + L_{tex}$ **(MAGnet)**	$\mathbf{25.2 \pm 1.2}$	$\mathbf{0.922 \pm 0.016}$

Table 4.5: Ablation Study on Loss Functions.

To demonstrate the necessity of the composite objective function defined in Equation 16, we trained the full MAGNet architecture with dual-path and attention using accumulating combinations of loss terms. This experimental design allows us to evaluate how much the specialized losses contribute to performance beyond the inherent structural benefits of the network itself.

Table4.5 summarizes the quantitative impact. The baseline configuration ($L_{adv} + L_{L1}$), essentially the objective of vanilla Pix2Pix [21], achieved a PSNR of 24.5 dB and an SSIM of 0.908 despite using the identical loss formulation. This strong performance floor is directly attributable to the underlying architectural design, which inherently minimizes errors even without specialized losses. Specifically, the high baseline SSIM indicates that the dual-Path strategy effectively decouples the geometry of bone and soft tissue. By solving the optimization conflict architecturally, the model achieves high structural fidelity using only the standard loss. Simultaneously, the high baseline PSNR suggests that the Mask-Guided Attention mechanism successfully filters out background noise during the encoding phase. Consequently, the auxiliary losses act as refinement mechanisms

rather than primary drivers. The addition of the L_{SSIM} yielded a focused improvement in geometric precision (+0.008 SSIM), stabilizing the high-frequency cortical edges where the pixel-wise loss may blur. Subsequently, the integration of L_{tex} provided the final boost in intensity accuracy (+0.7 dB), calibrating the fine-grained soft-tissue textures that the L_1 loss tends to smooth over.

Effectiveness of architecture

Config	Description	PSNR↑	SSIM↑
A	U-Net [74]	22.5 ± 1.8	0.851 ± 0.024
B	Dual-Branch U-Net (w/o Attention)	23.7 ± 1.8	0.882 ± 0.023
C	Attention U-Net	23.9 ± 1.4	0.906 ± 0.014
D	**MAGNet**	$\mathbf{25.2 \pm 1.2}$	$\mathbf{0.922 \pm 0.016}$

Table 4.6: Ablation Study on MAGNet architectural configurations.

We conducted a comprehensive factorial ablation study to quantify the contributions of the two core ideas proposed in our MAGnet. We evaluated four distinct configurations to isolate the impact of disentangling anatomical structures versus the impact of attentional feature filtering.

The experimental configurations are defined as follows. Config A is a single stream generator similar to the Pix2Pix architecture, directly mapping MRI to CT. B extends this by using two parallel paths without attention, isolating the contribution of anatomical splitting. C retains a single stream but integrates standard Attention Gates to evaluate whether learned implicit attention alone is sufficient. Finally, MAGNet combines the dual-branch design with mask-guided attention, representing the full proposed framework. The results of this ablation are summarized in Table 4.6.

Building on the architectural configurations described above, the experimental results reveal a clear progression in synthesis quality. The vanilla U-Net achieves a PSNR of 22.5 dB, which is slightly higher than the standard Pix2Pix baseline potentially due to the enhanced loss objective. However, it remains the weakest configuration, indicating that a single-stream model struggles to reconcile heterogeneous tissue characteristics within a shared representation. Introducing architectural separation through the Dual-Branch results in a measurable improvement of $+1.2$ dB in PSNR and $+0.031$ in SSIM. This gain shows that assigning separate pathways to bone and soft tissue helps

disentangle their distinct structural patterns and reduces cross-tissue interference.The Attention U-Net reaches 23.9 dB, which is slightly higher than the dual-branch model. This outcome demonstrates that standard attention gates provide substantial benefits even without explicit anatomical separation, since they suppress irrelevant background responses and stabilize feature learning. The full MAGNet framework achieves the highest performance at 25.2 dB. This score is significantly higher than both the Dual-Branch and Attention U-Net variants, confirming that mask-guided attention is complementary rather than redundant. By guiding the dual pathways using anatomical priors, instead of relying on unguided separation or implicit attention, MAGNet produces a synergistic effect that leads to a $+1.3$ dB improvement over the strongest ablation baseline.

Crucially, the success of MAGNet validates the decision to employ a learned soft-attention gate rather than a simple hard masking of features. While hard masking effectively isolates the Region of Interests (ROI), it zeroes out background gradients entirely, preventing the network from learning contextual cues at the immediate periphery of anatomical structures. By incorporating the mask as a bias term within a soft sigmoid activation, our approach retains gradient flow and allows the model to correct for minor inaccuracies in the automated TotalSegmentator priors. This ensures that the generated bone structures integrate naturally with the surrounding soft tissue rather than appearing as artificially pasted patches.

Effectiveness of Residual Refinement

Configuration	**PSNR (dB)** $\uparrow$	**SSIM** $\uparrow$
MAGNet (w/o Residual)	24.9 ± 1.3	0.914 ± 0.019
Full MAGNet	$\mathbf{25.2 \pm 1.2}$	$\mathbf{0.922 \pm 0.016}$

Table 4.7: Ablation Study on Residual Fusion in MAGNet.

Finally, we evaluated the contribution of the Residual Refinement Module ($\mathcal{R}$). As detailed in Section 3.2.5, the linear summation of the dual-branch outputs relies on binary masks to delineate tissue boundaries. Without further refinement, this hard fusion limits the model's ability to model the partial volume effects and smooth transitions naturally found in medical imaging.

To quantify this effect, we compared the MAGNet framework against a variant where the refinement module was ablated. Removing the residual module results in a performance drop of 0.3

dB in PSNR and 0.008 in SSIM. While the dual-branch architecture successfully generates accurate textures for bone and soft tissue independently, the drop in SSIM suggests that the hard nature of the binary mask priors introduces structural inconsistencies at the interface regions. The inclusion of the residual module recovers this fidelity. This quantitative improvement confirms that the module effectively acts as a learned blending operator, ensuring topological continuity between the rigid cortical bone and the deformable soft tissues without requiring manual intervention.

Chapter 5

Conclusions and Future Work

In this thesis, a novel anatomy-aware generative framework, MAGNet, has been developed for MRI-to-CT synthesis in the anatomically complex pelvic region. This chapter first presents the research contributions and conclusions derived from the experimental results, followed by a discussion of the study's limitations, and finally provides suggestions for future research directions.

5.1 Conclusion

The transition toward MRI-only radiotherapy offers a paradigm shift in cancer treatment, promising to reduce geometric uncertainties caused by registration errors while eliminating the ionizing radiation exposure associated with planning CTs. However, the safe clinical adoption of this workflow hinges on the ability to synthesize pseudo-CT images that are not only visually plausible but geometrically precise and texturally accurate.

The primary objective of MAGnet was to overcome the limitations of uniform translation models, which often fail to reconcile the conflicting statistical distributions of cortical bone and soft tissue. The principal contribution of this research is the development of an anatomy guided dual stream framework. Unlike standard U-Net or ResNet architectures that process the entire volume uniformly, this framework disentangles the generative task by allocating specialized model capacity to bone and soft tissue. This separation effectively resolved the hollow bone artifact observed in competing methods, ensuring that trabecular density is accurately preserved.

To improve feature selection within this dual-path topology, we integrated Mask-Guided Attention Gates into the generator. Distinct from standard attention mechanisms that rely on implicit, learned gating, this approach leverages automated priors from TotalSegmentator to explicitly enforce semantic constraints at every resolution level. This innovation allows the network to suppress background noise and strictly prevents information leakage between the bone and soft-tissue streams, ensuring that each branch focuses exclusively on its designated anatomical domain.

Furthermore, to ensure topological continuity, a Residual Refinement Module was introduced to fuse the dual-stream outputs. While binary masks provide strong guidance, they can inherently introduce high-frequency artifacts at tissue interfaces. The refinement module addresses this by predicting a residual correction map to smooth these transitions, ensuring seamless anatomical integration without altering the accurate HU values within the organs.

Finally, this work establishes the framework's robustness on heterogeneous Datasets. Through rigorous validation on both the standardized Gold Atlas benchmark and a heterogeneous internal clinical cohort (LMU Munich), we demonstrated that MAGNet consistently achieves superior synthesis quality. Crucially, the model maintains high performance (23.9 dB PSNR) even across diverse scanner vendors where standard baselines failed to converge, suggesting a high degree of clinical viability.

5.2 Limitations

While the research findings have revealed the effectiveness of the proposed MAGNet framework, several limitations in the current study must be acknowledged to contextualize its clinical applicability:

- MAGNet currently employs voxel-wise supervision, which prefers high-quality MRI-CT pairs for training. While this thesis has extensively demonstrated the geometric risks associated with unsupervised learning, specifically the tendency of models like CycleGAN to hallucinate anatomical structures, it is important to acknowledge the logistical trade-off. Unsupervised frameworks possess a specific advantage in data accessibility, as they can leverage vast archives of unpaired clinical scans. Our approach is inherently constrained to scenarios

where paired acquisition is standard practice, which may limit its immediate scalability in centers where multi-modality data is scarce.

- The current implementation operates on 2D axial slices. While computationally efficient, this approach ignores volumetric consistency along the z-axis, which can occasionally lead to minor artifacts in sagittal or coronal reconstructions, particularly in regions with rapid anatomical changes such as the femoral heads. This lack of inter slice context poses a potential challenge for extending the research to volumetric applications where 3D continuity is critical.

- The synthesis fidelity is intrinsically linked to the quality of the segmentation masks. Although the soft attention mechanism provides tolerance for minor errors, significant segmentation failures, such as the misclassification of bowel gas as bone, could propagate to the final CT synthesis, potentially leading to localized dosimetric inaccuracies.

- The proposed framework introduces computational overhead compared to standard end-to-end generators. The dependency on an external segmentation pipeline, TotalSegmentator for prior extraction, combined with the dual-path architecture, increases both the inference latency and the requirements for computational resources. This complexity may hinder real-time deployment in resource constrained clinical environments or on older radiotherapy console hardware.

5.3 Suggestions for Future Work

The findings of this thesis open several routes for further investigation to enhance the clinical utility of the proposed models. Future work could focus on the following directions:

- **Extended Validation on Internal Dataset:** A primary objective for future work is to renew data access agreements with the LMU Munich clinical partners. Regaining access to the internal dataset is critical to benchmark MAGNet against advanced architectures like ResViT and MaskGAN. This will verify whether anatomy-guided priors continue to offer superior stability compared to state-of-the-art models.

- **Dosimetric Verification**: While image quality metrics serve as strong proxies, the ultimate validation for radiotherapy is dosimetric accuracy. Future studies should involve importing synthesized CTs into clinical systems to calculate photon and proton dose distributions. This process must incorporate qualitative assessment by clinical experts, such as radiation oncologists and medical physicists, to verify the clinical acceptability of the generated plans and Dose-Volume Histograms (DVH) against those derived from ground truth CTs.

- **Multi-Modality Integration:** The current framework relies solely on T2-weighted MRI in the axial plane. Incorporating multi-contrast inputs or multi-view data could provide complementary information, helping to resolve ambiguities between air pockets and cortical bone that remain challenging for single-sequence models.

- **Robustness to Pathology and Implants**: The present work focuses on anatomically typical cases and excludes patients with metallic implants or severe artifacts to ensure stable training and evaluation. Future research should extend MAGNet to cohorts with hip prostheses, post-surgical alterations, and pathological bone lesions, systematically characterizing failure modes and adapting the model or anatomical priors to handle metal-induced distortions and atypical anatomy. Establishing robustness in these challenging scenarios is essential for translating MRI-only workflows to the full spectrum of patients encountered in routine clinical practice.

Bibliography

[1] CJ Dean, JR Sykes, RA Cooper, P Hatfield, B Carey, S Swift, SE Bacon, D Thwaites, D Sebag-Montefiore, and AM Morgan. An evaluation of four ct–mri co-registration techniques for radiotherapy treatment planning of prone rectal cancer patients. *The British journal of radiology*, 85(1009):61–68, 2012.

[2] Alankrita Aggarwal, Mamta Mittal, and Gopi Battineni. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, 1(1):100004, 2021.

[3] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry L Prince, and Zongben Xu. Unsupervised mr-to-ct synthesis using structure-constrained cyclegan. *IEEE transactions on medical imaging*, 39(12):4249–4261, 2020.

[4] Tugba Akinci D'Antonoli, Lucas K Berger, Ashraya K Indrakanti, Nathan Vishwanathan, Jakob Weiss, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, et al. Totalsegmentator MRI: Robust sequence-independent segmentation of multiple anatomic structures in MRI. *Radiology*, 314(2):e241613, 2025.

[5] Rajamanickam Baskar, Kuo Ann Lee, Richard Yeo, and Kheng-Wei Yeoh. Cancer and radiation therapy: current advances and future directions. *International journal of medical sciences*, 9(3):193, 2012.

[6] Krishna Koka, Amit Verma, Bilikere S Dwarakanath, and Rao VL Papineni. Technological advancements in external beam radiation therapy (ebrt): An indispensable tool for cancer treatment. *Cancer Management and Research*, pages 1421–1429, 2022.

[7] Sibel Karaca and Meltem Kırlı Bölükbaş. Time matters: A review of current radiotherapy practices and efficiency strategies. *Technology in Cancer Research & Treatment*, 24:15330338251345376, 2025.

[8] Wouter van Elmpt and Guillaume Landry. Quantitative computed tomography in radiation therapy: A mature technology with a bright future, 2018.

[9] Sonja Stieb, Brigid McDonald, Mary Gronberg, Grete May Engeseth, Renjie He, and Clifton David Fuller. Imaging for target delineation and treatment planning in radiation oncology: current and emerging techniques. *Hematology/oncology clinics of North America*, 33(6):963–975, 2019.

[10] Maria A Schmidt and Geoffrey S Payne. Radiotherapy planning using mri. *Physics in Medicine & Biology*, 60(22):R323, 2015.

[11] Hayder Alabedi. Assessing setup errors and shifting margins for planning target volume in head, neck, and breast cancer. *Journal of Medicine and Life*, 16(3):394, 2023.

[12] Hussain Almohiy. Paediatric computed tomography radiation dose: a review of the global dilemma. *World journal of radiology*, 6(1):1, 2014.

[13] Amir M Owrangi, Peter B Greer, and Carri K Glide-Hurst. MRI-only treatment planning: benefits and challenges. *Physics in Medicine & Biology*, 63(5):05TR01, 2018.

[14] Yuan Gao, Chih-Wei Chang, Sagar Mandava, Raanan Marants, Jessica E Scholey, Matthew Goette, Yang Lei, Hui Mao, Jeffrey D Bradley, Tian Liu, et al. Mri-only based material mass density and relative stopping power estimation via deep learning for proton therapy: a preliminary study. *Scientific Reports*, 14(1):11166, 2024.

[15] Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of MRI, CT and PET: Review and analysis. *Medical image analysis*, 92:103046, 2024.

[16] Marion Boulanger, Jean-Claude Nunes, Hilda Chourak, Axel Largent, Safaa Tahri, Oscar Acosta, R De Crevoisier, Caroline Lafond, and Anais Barateau. Deep learning methods to

generate synthetic ct from mri in radiotherapy: a literature review. *Physica Medica*, 89:265–281, 2021.

[17] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Zongben Xu, and Jerry Prince. Unpaired brain mr-to-ct synthesis using a structure-constrained cyclegan. In *International Workshop on Deep Learning in Medical Image Analysis*, pages 174–182. Springer, 2018.

[18] Jelmer M Wolterink, Tim Leiner, Max A Viergever, and Ivana Išgum. Generative adversarial networks for noise reduction in low-dose ct. *IEEE transactions on medical imaging*, 36(12):2536–2545, 2017.

[19] Vu Minh Hieu Phan, Zhibin Liao, Johan W Verjans, and Minh-Son To. Structure-preserving synthesis: Maskgan for unpaired mr-ct translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 56–65. Springer, 2023.

[20] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1125–1134, 2017.

[22] R Schofield, L King, U Tayal, I Castellano, J Stirrup, F Pontana, James Earls, and E Nicol. Image reconstruction: Part 1–understanding filtered back projection, noise and image acquisition. *Journal of cardiovascular computed tomography*, 14(3):219–225, 2020.

[23] Michael Chappell. *Principles of Medical Imaging for Engineers*. Springer, 2019.

[24] Rolf Pohmann. Physical basics of nmr. In *In vivo NMR Imaging: Methods and Protocols*, pages 3–21. Springer, 2011.

[25] Philip Mayles, Alan Nahum, and Jean-Claude Rosenwald. *Handbook of radiotherapy physics: theory and practice*. CRC Press, 2007.

[26] Emily Johnstone, Jonathan J Wyatt, Ann M Henry, Susan C Short, David Sebag-Montefiore, Louise Murray, Charles G Kelly, Hazel M McCallum, and Richard Speight. Systematic review of synthetic computed tomography generation methodologies for use in magnetic resonance imaging–only radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 100(1):199–217, 2018.

[27] R Prabhakar, PK Julka, T Ganesh, A Munshi, RC Joshi, and GK Rath. Feasibility of using mri alone for 3d radiation treatment planning in brain tumors. *Japanese journal of clinical oncology*, 37(6):405–411, 2007.

[28] David Pasquier, Nacim Betrouni, Maximilien Vermandel, Thomas Lacornerie, Eric Lartigau, and Jean Rousseau. Mri alone simulation for conformal radiation therapy of prostate cancer: technical aspects. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 160–163. IEEE, 2006.

[29] DC Weber, H Wang, S Albrecht, M Ozsahin, E Tkachuk, M Rouzaud, P Nouet, and G Dipasquale. Open low-field magnetic resonance imaging for target definition, dose calculations and set-up verification during three-dimensional crt for glioblastoma multiforme. *Clinical Oncology*, 20(2):157–167, 2008.

[30] Anna M Dinkla, Rob van der Laarse, Emmie Kaljouw, Bradley R Pieters, Kees Koedooder, Niek van Wieringen, and Arjan Bel. A comparison of inverse optimization algorithms for hdr/pdr prostate brachytherapy treatment planning. *Brachytherapy*, 14(2):279–288, 2015.

[31] Jason A Dowling, Jidi Sun, Peter Pichler, David Rivest-Hénault, Soumya Ghose, Haylea Richardson, Chris Wratten, Jarad Martin, Jameen Arm, Leah Best, et al. 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*, 93(5):1144–1153, 2015.

[32] Adam Johansson, Anders Garpebring, Mikael Karlsson, Thomas Asklund, and Tufve Nyholm. Improved quality of computed tomography substitute derived from magnetic resonance (mr) data by incorporation of spatial information–potential application for mr-only radiotherapy and

attenuation correction in positron emission tomography. *Acta oncologica*, 52(7):1369–1373, 2013.

[33] Ninon Burgos, M Jorge Cardoso, Filipa Guerreiro, Catarina Veiga, Marc Modat, Jamie McClelland, Antje-Christin Knopf, Shonit Punwani, David Atkinson, Simon R Arridge, et al. Robust ct synthesis for radiotherapy planning: application to the head and neck region. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 476–484. Springer, 2015.

[34] Daniel Andreasen, Koen Van Leemput, Rasmus H Hansen, Jon AL Andersen, and Jens M Edmund. Patch-based generation of a pseudo ct from conventional mri sequences for mri-only radiotherapy of the brain. *Medical physics*, 42(4):1596–1605, 2015.

[35] Hossein Arabi and Habib Zaidi. Magnetic resonance imaging-guided attenuation correction in whole-body pet/mri using a sorted atlas approach. *Medical image analysis*, 31:1–15, 2016.

[36] Xiao Han. Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical physics*, 44(4):1408–1419, 2017.

[37] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International conference on medical image computing and computer-assisted intervention*, pages 417–425. Springer, 2017.

[38] Hajar Emami, Ming Dong, Siamak P Nejad-Davarani, and Carri K Glide-Hurst. Generating synthetic cts from magnetic resonance images using generative adversarial networks. *Medical physics*, 45(8):3627–3636, 2018.

[39] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[40] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[41] Jelmer M. Wolterink, Anna M. Dinkla, Mark H. F. Savenije, Peter R. Seevinck, Cornelis A. T. van den Berg, and Ivana Išgum. Deep mr to ct synthesis using unpaired data. In Sotirios A. Tsaftaris, Ali Gooya, Alejandro F. Frangi, and Jerry L. Prince, editors, *Simulation and Synthesis in Medical Imaging*, pages 14–23, Cham, 2017. Springer International Publishing.

[42] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pages 529–536. Springer, 2018.

[43] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[44] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[45] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[46] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[48] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[49] Yuxin Hu, Han Zhou, Ning Cao, Can Li, and Can Hu. Synthetic ct generation based on cbct using improved vision transformer cyclegan. *Scientific Reports*, 14(1):11455, 2024.

[50] Shaoyan Pan, Elham Abouei, Jacob Wynne, Chih-Wei Chang, Tonghe Wang, Richard LJ Qiu, Yuheng Li, Junbo Peng, Justin Roper, Pretesh Patel, et al. Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548, 2024.

[51] Mohamed A Bahloul, Saima Jabeen, Sara Benoumhani, Habib Abdulmohsen Alsaleh, Zehor Belkhatir, and Areej 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*, 25(11):e14499, 2024.

[52] Uulke A van der Heide, Marloes Frantzen-Steneker, Eleftheria Astreinidou, Marlies E Nowee, and Petra J van Houdt. MRI basics for radiation oncologists. *Clinical and translational radiation oncology*, 18:74–79, 2019.

[53] Edward L Nickoloff and Philip O Alderson. Radiation exposures to patients from CT: reality, public perception, and policy. *American Journal of Roentgenology*, 177(2):285–287, 2001.

[54] Yi-Qiu Zhang, Peng-Cheng Hu, Run-Ze Wu, Yu-Shen Gu, Shu-Guang Chen, Hao-Jun Yu, Xiang-Qing Wang, Jun Song, and Hong-Cheng Shi. The image quality, lesion detectability, and acquisition time of 18F-FDG total-body PET/CT in oncological patients. *European journal of nuclear medicine and molecular imaging*, 47(11):2507–2515, 2020.

[55] Paul Kyu Han, Debra E Horng, Kuang Gong, Yoann Petibon, Kyungsang Kim, Quanzheng Li, Keith A Johnson, Georges El Fakhri, Jinsong Ouyang, and Chao Ma. MR-based PET attenuation correction using a combined ultrashort echo time/multi-echo Dixon acquisition. *Medical physics*, 47(7):3064–3077, 2020.

[56] Xuanru Zhou, Wenwen Cai, Jiajun Cai, Fan Xiao, Mengke Qi, Jiawen Liu, Linghong Zhou, Yongbao Li, and Ting Song. Multimodality MRI synchronous construction based deep learning framework for MRI-guided radiotherapy synthetic CT generation. *Computers in Biology and Medicine*, 162:107054, 2023.

[57] Jens Sjölund, Daniel Forsberg, Mats Andersson, and Hans Knutsson. Generating patient specific pseudo-CT of the head from MR using atlas-based regression. *Physics in Medicine & Biology*, 60(2):825, 2015.

[58] Mateusz C Florkow, Koen Willemsen, Vasco V Mascarenhas, Edwin HG Oei, Marijn van Stralen, and Peter R Seevinck. Magnetic resonance imaging versus computed tomography for three-dimensional bone imaging of musculoskeletal pathologies: a review. *Journal of Magnetic Resonance Imaging*, 56(1):11–34, 2022.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 770–778, 2016.

[60] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[61] Seung Kwan Kang, Hyun Joon An, Hyeongmin Jin, Jung-in Kim, Eui Kyu Chie, Jong Min Park, and Jae Sung Lee. Synthetic CT generation from weakly paired MR images using cycle-consistent GAN for MR-guided radiotherapy. *Biomedical engineering letters*, 11(3):263–271, 2021.

[62] Kishore Krishnagiri Manoj Doss and Jyh-Cheng Chen. Utilizing deep learning techniques to improve image quality and noise reduction in preclinical low-dose PET images in the sinogram domain. *Medical Physics*, 51(1):209–223, 2024.

[63] Changfei Gong, Junming Jian, Yuling Huang, Mingming Luo, Shenggou Ding, Xingxing Yuan, Xiaoping Wang, and Yun Zhang. Boundary information-guided adversarial diffusion model for efficient unsupervised synthetic CT generation. *Medical Physics*, 52(6):4675—-4693, 2025.

[64] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

[65] Azin Shokraei Fard, David C Reutens, and Viktor Vegh. From CNNs to GANs for cross-modality medical image estimation. *Computers in biology and medicine*, 146:105556, 2022.

[66] Jiaxu Zheng, Zhenrong Shen, Lichi Zhang, and Qun Chen. Structure-guided MR-to-CT synthesis with spatial and semantic alignments for attenuation correction of whole-body PET/MR imaging. *Medical Image Analysis*, 103:103622, 2025.

[67] Si Young Yie, Siyeop Yoon, Jaewon Yang, Kyungsang Kim, Jae Sung Lee, and Quanzheng Li. Score-based-diffusion-model-based synthetic CT generation from MR images and post-hoc uncertainty analysis. In *Medical Imaging 2025: Image Processing*, volume 13406, pages 600–604. SPIE, 2025.

[68] Yu Luo, Shaowei Zhang, Jie Ling, Zhiyi Lin, Zongming Wang, and Shun Yao. Mask-guided generative adversarial network for MRI-based CT synthesis. *Knowledge-Based Systems*, 295:111799, 2024.

[69] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. MR and CT data with multiobserver delineations of organs in the pelvic area—part of the Gold Atlas project. *Medical physics*, 45(3):1295–1300, 2018.

[70] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.

[71] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 2223–2232, 2017.

[72] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

[73] Jonas Denck, Jens Guehring, Andreas Maier, and Eva Rothgang. Mr-contrast-aware image-to-image translations with generative adversarial networks. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2069–2078, 2021.

[74] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.