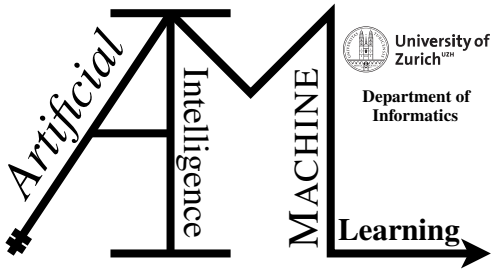
Fullbody Synthetic CT Generation using Deep Learning Methods

Master Project

Flavian Roland Thür, Mike Jason Frei, Nicolas Yann Schuler
16-562-274, 20-488-144, 18-618-454

Submitted on
April 29, 2026

Thesis Supervisor
Prof. Dr. Manuel Günther
Mariia Lapaeva, PD Dr. sc. nat. Stephanie Tanadini-Lang, Dr. rer. nat. Riccardo Dal Bello



Master Project

Author: Flavian Roland Thür, Mike Jason Frei, Nicolas Yann Schuler

Project period: July 17, 2025 - April 29, 2026

Artificial Intelligence and Machine Learning Group
Department of Informatics, University of Zurich

Acknowledgements

Acknowledgements

We would like to express our sincere gratitude to everyone who supported and guided us throughout this project.

First and foremost, we thank Prof. Dr. Manuel Günther for supervising this project, for his continued feedback and guidance, and for providing us with the opportunity to work on this project.

We owe particular thanks to Mariia Lapaeva, whose consistent and dedicated supervision shaped the direction of this project at every stage. Her deep expertise and thoughtful guidance ensured the project stayed on track, while her motivating presence made the collaboration genuinely enjoyable.

We are also grateful to PD Dr. sc. nat. Stephanie Tanadini-Lang and Dr. rer. nat. Riccardo Dal Bello for their constructive feedback on our results and their support in discussing key research questions throughout the project.

Special thanks go to Philipp Wallimann for his assistance with the N-Peaks normalization method, whose insights were directly relevant to one of the central components of this work.

Finally, we thank Nico Zala, whose prior work on this project provided a valuable foundation and whose practical help with code and implementation saved us considerable time.

Abstract

Synthetic computed tomography (sCT) generation from magnetic resonance images (MRI) is a key enabler of MR-only radiotherapy, eliminating the need for a separate computed tomography (CT) acquisition while preserving the soft-tissue contrast advantages of MRI. This project investigates GAN-based MR-to-CT image translation in a multi-center, multi-anatomical setting using data from the SynthRAD2023 and 2025 challenges. Three methodological dimensions are systematically examined: the trade-off between joint fullbody and region-specific training, the effect of MRI intensity normalization strategies on translation quality, and the influence of generator architecture and training regime. Experiments are conducted across three 2D image translation frameworks – Pix2Pix, CUT, and CycleGAN – and evaluated using image similarity metrics as well as a dosimetric assessment. The results show that the choice of translation framework has a larger influence on performance than any individual preprocessing or architectural decision, with Pix2Pix and CUT consistently outperforming CycleGAN across conditions. Among the tested generator backbones, a ResNet-9 architecture achieved the best image similarity scores, and adversarial training proved essential for structural coherence. These findings provide a structured account of design trade-offs in multi-center MR-to-CT translation and offer practical guidance for future work toward clinically deployable MR-only radiotherapy workflows.

Contents

Chapter 1

Introduction

Radiotherapy is used in the treatment of more than half of all cancer patients (Thummerer et al., 2024). Delivering radiation accurately requires precise knowledge of tissue photon attenuation throughout the patient volume, as treatment planning systems rely on this information to model radiation transport and optimize dose delivery. Computed tomography (CT) is the established standard for this purpose: its Hounsfield unit (HU) values provide quantitative electron density maps that planning systems use directly (Owrangi et al., 2018). Magnetic resonance imaging (MRI), by contrast, offers superior soft-tissue contrast without ionizing radiation, making it increasingly indispensable for tumor delineation, organ-at-risk contouring, and real-time treatment guidance (Schmidt and Payne, 2015). However, MRI signal intensity reflects proton density and relaxation properties rather than electron density, and cannot be directly converted to the attenuation information required for dose calculation (Jonsson et al., 2010).

Conventional radiotherapy workflows therefore rely on both modalities: MRI for delineation, CT for dose calculation. When acquired in separate sessions – often days apart – the two scans are subject to anatomical inconsistencies arising from patient repositioning, organ motion, bowel filling, or weight changes, all of which complicate inter-modality image registration and can introduce systematic errors into the treatment plan (Owrangi et al., 2018). While MR-guided radiotherapy (MRgRT) partially addresses this by enabling real-time MRI during treatment delivery, CT acquisition for initial planning remains necessary even in hybrid MRgRT systems. MR-only radiotherapy represents a more fundamental shift: by generating a *synthetic CT* (sCT) directly from the MRI, the workflow is consolidated into a single imaging session, eliminating registration uncertainty entirely, reducing patient burden, and removing additional radiation exposure. Beyond workflow simplification, this approach is particularly valuable for adaptive radiotherapy, where repeated imaging is required to account for anatomical changes over the course of treatment – making the elimination of repeated CT acquisitions both dosimetrically and logistically substantial. MR-only radiotherapy is already clinically established for the prostate and increasingly feasible for sites such as the brain and head-and-neck (Owrangi et al., 2018).

Deep learning, and specifically image-to-image translation, has become the dominant paradigm for sCT generation (Boulanger et al., 2021). In this setting, a generator model serves as a mapping function transforming an input MRI into its corresponding sCT. Conditional generative adversarial networks (cGAN) such as Pix2Pix (Isola et al., 2017) have demonstrated strong performance under paired training, while unpaired frameworks such as CycleGAN (Zhu et al., 2017) and contrastive unpaired translation (CUT) (Park et al., 2020) have extended applicability to settings where aligned MRI – CT pairs are unavailable. Despite this progress, most published work trains and evaluates models within a single anatomical region and on data from a homogeneous scanner environment. Translating these results into a clinical deployment setting raises three practical challenges that this project addresses directly.

The first challenge concerns *anatomical generalization*. In clinical practice, a radiotherapy de-

partment treats patients across a range of anatomical sites. Maintaining separate, site-specific models for each region is operationally burdensome. A single full-body model trained jointly on all anatomical regions is therefore an attractive deployment target – but it is not clear whether such a model can match the accuracy of region-specific models that benefit from anatomically homogeneous training data. Each region presents distinct imaging characteristics, tissue compositions, and dose-relevant structures; a joint model must accommodate all of these within a shared parameter space. Whether this accommodation incurs a meaningful accuracy cost is an open empirical question.

The second challenge concerns *multi-center scanner heterogeneity*. MRI signal intensity is inherently non-standardized: unlike CT's HU scale, MRI intensities depend on field strength, coil configuration, acquisition sequence, and scanner manufacturer. When training data originate from multiple centers models must learn mappings from intensities that carry scanner-specific rather than purely tissue-specific information. Whether tissue-based normalization methods, which harmonize intensity distributions by aligning tissue-derived histogram landmarks, can outperform simpler global rescaling approaches in this setting remains an open question. As an alternative to preprocessing-based harmonization, a model-based approach can be pursued by conditioning the network on scanner identity through a dedicated input layer per acquisition center, effectively shifting the burden of domain adaptation from the data to the model architecture.

The third challenge concerns *framework and architecture design*. While multiple image-to-image translation frameworks have been applied to sCT generation, direct comparisons under controlled conditions are rare, as studies differ in dataset, preprocessing, and evaluation protocol. Within a given framework such as Pix2Pix, it further remains unclear how much of the performance is attributable to the adversarial training objective versus the generator architecture itself – and whether a well-designed generator trained without a discriminator can achieve comparable results. Disentangling these factors is important for understanding what drives sCT quality and for making principled architecture choices in future work.

Motivated by these three challenges, this project systematically investigates sCT generation using data from the SynthRAD2023 and SynthRAD2025 challenges (Huijben et al., 2024; Thummerer et al., 2024), which provide paired MRI – CT volumes across multiple anatomical regions and acquisition centers. Three research questions structure the investigation:

- **RQ1:** Can a jointly trained fullbody model achieve comparable performance to region-specific models for MR-to-CT translation, under otherwise identical training conditions?
- **RQ2:** How do different MRI intensity normalization techniques affect sCT quality, and can tissue-based normalization improve performance over global approaches in the presence of multi-center scanner heterogeneity?
- **RQ3:** How does the choice of generator architecture and training regime influence sCT quality within the Pix2Pix framework, and how does adversarial training compare to a generator-only setting?

To address these questions, the remainder of this report is structured as follows. Chapter 2 surveys prior work on sCT generation and image-to-image translation. Chapter 3 introduces the methodological foundations. Chapter 4 describes the datasets, and Chapter 5 details the preprocessing pipeline and experimental building blocks. Chapter 6 specifies the experimental design and presents results. Chapter 7 examines the findings critically and identifies directions for future work. Chapter 8 synthesizes the main conclusions.

Chapter 2

Related Work

Different approaches have been developed to generate sCT images from MRI data. These can be broadly grouped into bulk density-based, atlas-based, and deep learning-based methods (Boulanger et al., 2021).

Bulk density-based methods segment MRI scans into a few tissue classes, typically air, soft tissue, and bone. Each segmented region is then assigned a fixed electron density value, which is used to calculate the dose distribution. Although simple in concept, this approach ignores variations within tissues, which can reduce accuracy. Depending on the implementation it can heavily depend on manual input and hence be time-consuming (Boulanger et al., 2021).

Atlas-based methods work by matching an MRI image to one or more example MRI – CT image pairs, called atlases. The target MRI is aligned (registered) with these reference images and then the deformation fields from registration are applied to the atlas CTs to warp them into the target's space. However, this method often struggles when the patient's anatomy differs from the atlases and requires substantial computing time due to multiple complex registrations (Boulanger et al., 2021).

Deep learning methods use models made up of many processing layers that automatically learn features from data. For sCT generation, deep learning models learn the relationship between MRI intensities and CT HU values. Once trained, they can generate an sCT from a new MRI. These methods are fast and typically do not require complex registration between different patients, as they can operate on individual patient data (Boulanger et al., 2021).

This chapter surveys prior work relevant to deep learning methods in the context of MRI-to-CT synthesis. It begins by reviewing the main image-to-image translation frameworks, followed by a discussion of data dimensionality choices. It then provides a broader empirical overview of deep learning methods for sCT generation. The chapter further covers fullbody approaches, and concludes with an overview of the SynthRAD2023 and SynthRAD2025 challenges, contextualizing the approaches explored in this report.

2.1 Image-To-Image Translation Frameworks

Deep learning architectures for sCT generation from MRI can broadly be categorized into generator-only approaches and generative adversarial networks (GANs), including their variants such as cGANs, least-squares GANs, and cycleGANs (Boulanger et al., 2021). With the recent rise of transformer-based and diffusion-based approaches, the field has considerably expanded, especially in the space of generator networks. These newer architectures can capture long-range dependencies and complex spatial relationships more effectively, enabling the generation of more realistic and detailed sCT images (Dayarathna et al., 2024).

2.1.1 Generator-only

For image-based tasks, **convolutional neural networks (CNN)** introduced by Lecun et al. (1998) are among the most widely used deep learning architectures. CNNs rely on convolutional filters (kernels) to automatically detect and learn image features. Within the convolutional layers, convolutions are applied using trainable filters, followed by nonlinear activation functions, which enable the model's ability to capture complex patterns in the data.

In the context of sCT generation from MRI, the generator model serves as a mapping function that transforms an input MRI into its corresponding sCT. During training, the generator minimizes a loss function that measures the intensity-based similarity between the generated sCT and the reference CT image. Most generator architectures used in sCT generation are based on convolutional encoder–decoder (CED) networks. The CED network consists of two main parts: an encoder, which progressively compresses the input image into higher-level feature representations, and a decoder, which reconstructs the sCT by increasing the spatial resolutions of these features using transposed convolutions. This structure enables the model to learn both global context and fine local details (Boulanger et al., 2021).

A widely adopted and highly effective CED variant is the UNet, originally proposed for biomedical image segmentation by Ronneberger et al. (2015). It extended the encoder–decoder design with skip connections that directly link corresponding layers between the encoder and decoder, allowing the network to recover spatial details lost during downsampling.

More recently, Transformer-based architectures have emerged as a powerful alternative to CNNs for sCT generation. Originally developed for sequence modeling in natural language processing, the Transformer architecture by Vaswani et al. (2017) has become a cornerstone of modern deep learning. Its key innovation lay in the self-attention mechanism, which enabled the model to capture long-range dependencies within the data more effectively than traditional convolutional or recurrent architectures.

In computer vision, the **Vision Transformer (ViT)** extended this concept by treating an image as a sequence of smaller, non-overlapping patches (Dosovitskiy et al., 2020). Each patch is embedded and processed through an encoder that integrates both visual and positional information, followed by a task-specific decoder. Within the encoder, multi-head self-attention (MSA) layers compute attention maps across all image patches, allowing the network to focus selectively on the most relevant regions. Thanks to this ability to model global image context, ViTs have achieved strong performance across various vision tasks such as classification, object detection, and segmentation. In recent years, Vision Transformer-based models have also been increasingly applied in medical imaging, demonstrating promising results in areas like organ segmentation, lesion detection, and cross-modality image synthesis (Shamshad et al., 2023). Their capacity to capture complex spatial relationships makes them particularly well suited for tasks like sCT generation from MRI, where understanding global anatomical structure is critical (Dayarathna et al., 2024).

The **Shifted Window (Swin) Transformer** (Liu et al., 2021) introduced a hierarchical vision transformer architecture designed to scale efficiently to high-resolution images while retaining strong inductive biases for dense prediction tasks. Instead of global self-attention, Swin partitions the input into non-overlapping local windows and applies window-based multi-head self-attention (W-MSA). Cross-window information exchange was enabled through a shifted window mechanism (SW-MSA) applied in alternating layers. This design, as depicted in Figure 2.1, yields linear computational complexity Ω with respect to image size $h \times w$ (compared to vision transformer's window-less MSA with quadratic complexity).

Merging patches increased the receptive field through concatenating neighboring patches and applying a linear layer. Pre-trained Swin Transformer of different sizes (Tiny, Small, Base, Large) have been trained on the natural image dataset ImageNet with input resolutions of 224^2 or 384^2. Originally, Swin achieved state-of-the-art results on image classification, object detection, and semantic segmentation in natural image benchmarks.

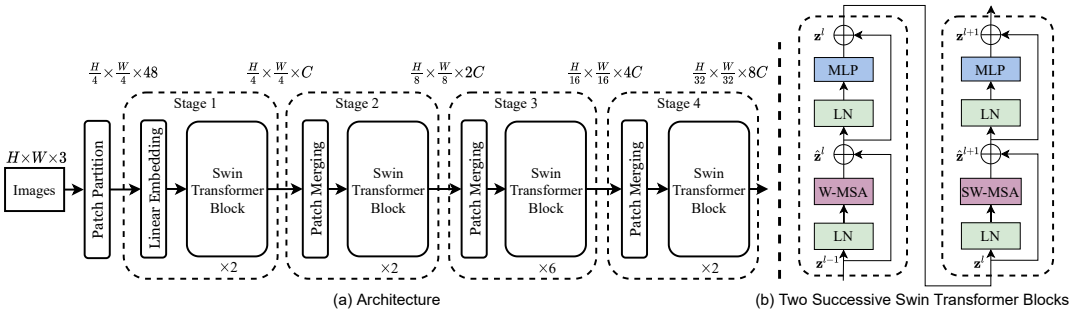


Figure 2.1: SWIN TRANSFORMER ARCHITECTURE. Swin Transformer-T V1 model architecture (a) with a detailed view of two consecutive Swin Transformer blocks (b) (Liu et al., 2021).

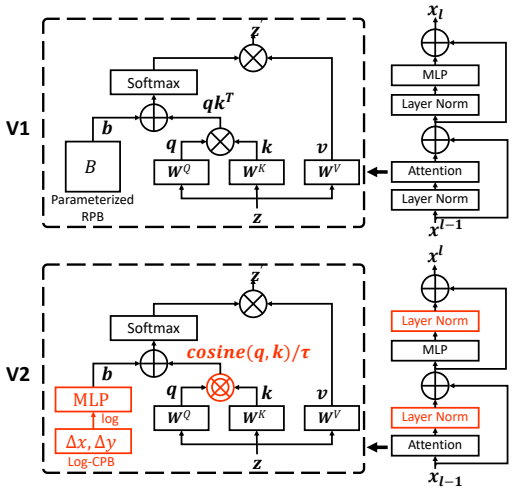


Figure 2.2: COMPARISION OF SWIN TRANSFORMER VERSIONS. Difference of Swin Transformer blocks in V1 and V2 (Liu et al., 2022).

Swin Transformer V2 (Liu et al., 2022) extended the original design to improve training stability and scalability, particularly for very large models and high-resolution inputs. Two key modifications, as shown in Figure 2.2, were introduced: (i) scaled cosine attention, which normalizes attention logits to prevent numerical instability when scaling model size, and (ii) log-spaced continuous relative position bias, enabling better generalization across varying image resolutions. Swin V2 supports significantly larger pre-training resolutions (up to 1536^2) and parameter counts, while preserving the hierarchical window-based structure of Swin. Pre-trained Swin Transformer of different sizes (Tiny, Small, Base, Large) have been trained on the natural image dataset ImageNet with input resolutions of 256^2 or 384^2.

Swin-UNet (Cao et al., 2023) adapted the Swin Transformer architecture to the medical image segmentation domain by integrating it into a U-shaped encoder–decoder framework with skip connections inspired by UNet Ronneberger et al. (2015). In contrast to hybrid CNN–Transformer models, Swin-UNet is a pure transformer architecture in which both encoder and decoder are composed of Swin Transformer blocks. Medical images are treated as 2D inputs. The decoder mirrors the encoder structure and uses patch expanding layers to progressively restore spatial resolution while reducing embedding dimensionality. Swin-UNet was evaluated on standard medical

segmentation benchmarks, including multi-organ CT and cardiac MRI datasets, demonstrating improved global context modeling and competitive performance compared to CNN-based and hybrid approaches. Similar to Swin, different model scales (e.g., Tiny and Base) were explored, showing that smaller variants already achieve strong performance, which is particularly relevant in data-limited medical settings.

Denoising diffusion probabilistic models (DDPMs), first introduced by Ho et al. (2020), represent a novel class of generative models that have shown remarkable success in producing high-quality and realistic images. These models learn to generate data through a gradual denoising process, starting from pure noise and progressively refining it into a coherent image across a sequence of time steps. The diffusion process consists of two stages: a forward process, where random Gaussian noise is iteratively added to an input image until it becomes nearly indistinguishable from random noise, and a reverse process, where the model learns to invert this transformation step by step. This reverse diffusion process effectively reconstructs the original data distribution by learning how to remove the noise added at each stage (Ho et al., 2020).

By training over many such transitions, DDPMs learn a powerful representation of the underlying data distribution, allowing them to generate highly realistic samples. However, despite their impressive performance, diffusion models are computationally intensive, as the generative process involves a large number of iterative denoising steps. Recent advancements, such as the latent diffusion model (LDM) proposed by Rombach et al. (2021), have significantly reduced this computational burden by performing the diffusion process in a compressed latent space rather than in the full image domain.

2.1.2 Generative Adversarial Networks

The generative adversarial network (GAN) framework was introduced by Goodfellow et al. (2020) as a general approach to generative modeling. In this approach, two neural networks, the generator and the discriminator, are trained simultaneously in a two-player game. The generator learns to produce sCT images that resemble real CT scans, while the discriminator learns to distinguish between real and generated images (Yi et al., 2019).

Unlike generator-only models, GANs introduce an additional adversarial loss, which acts as a data-driven regularizer that encourages the generated images to match the true data distribution. In the original formulation, both the generator and discriminator were implemented as multilayer perceptrons (MLP), but modern GANs typically use CNNs for both components. The generator is often based on a UNet or ResNet architecture, as these networks are effective in capturing fine structural details while maintaining global consistency (Boulanger et al., 2021).

In the review by Boulanger et al. (2021), GAN-based architectures accounted for 55% of the reviewed studies and spanned three main variants: classical GAN, cGAN, and cycle-GAN. Among these, cGAN was the most widely used, appearing in 20 studies and often being implemented as Pix2Pix. In contrast, cycle-GAN offered the notable advantage of not requiring paired MRI/CT data for training. Both architectures are discussed in greater detail in Sections 3.3.1 and 3.3.2, respectively.

2.2 Data Dimensionality (2D / 2.5D / 3D)

In deep learning-based sCT generation, models differ in how they process volumetric image data. *2D models* operate on individual image slices (e.g. axial, sagittal, or coronal), while *3D models* process volumetric patches that capture spatial context across all three dimensions. *2.5D approaches* represent a middle ground, typically by stacking multiple neighboring 2D slices as input channels to approximate 3D context without the full computational cost of volumetric processing. When

examining the configurations of deep learning models for sCT generation, Spadea et al. (2021) found that most studies have employed 2D architectures, followed by 3D patch-based and 2.5D approaches.

The comparison between 2D and 3D models has produced mixed findings. While Fu et al. (2019) reported that modified 3D UNet architectures outperform their 2D counterparts, Neppl et al. (2019) found the opposite, showing that 2D models can achieve higher image similarity and better dosimetric accuracy in certain settings. Such discrepancies are often attributed to differences in patch size, dataset quality, and model depth, which strongly influence the ability of 3D networks to capture contextual information. Moreover, because 3D models are computationally more demanding, requiring more memory and training time, they may lose global spatial context when restricted to smaller patch sizes (Spadea et al., 2021). As a practical compromise, Oulbacha and Kadoury (2020) proposed 2.5D consisting of stacking multiple neighboring 2D slices as input channels, approximating 3D spatial context at a fraction of the computational cost (Dayarathna et al., 2024).

2.3 Deep Learning for sCT Generation

Building on the architectural families and pretrained models discussed in the preceding sections, this section synthesizes the broader empirical findings from the literature on deep learning-based sCT generation, providing a quantitative and comparative perspective on how different model families have been adopted and evaluated across the field.

2.3.1 Literature Review: Achievements

In the systematic review by Spadea et al. (2021), covering publications from January 2014 to December 2020, GAN-based architectures accounted for the majority of sCT generation studies (55%), followed by UNet-based models (35%) and other CNN variants (10%). Similarly, Boulanger et al. (2021) reported that around 45% of 57 reviewed sCT generation studies employed generator-only architectures including UNet, fully convolutional networks, ResNet, SE-ResNet, and DenseNet.

While this distribution reflects a clear preference for adversarial training during period reviewed by Spadea et al. (2021), it does not imply architectural superiority: comparative studies have reported mixed outcomes, with some finding GANs outperforming UNet or other CNN-based models, while others observed similar or even inferior performance. These inconsistencies suggest that model architecture alone does not determine success. As highlighted by Largent et al. (2020), the loss function (i.e. training objective) plays a crucial role in guiding network learning and can strongly influence final image quality – yet in many studies it has received considerably less attention than architectural design, despite its fundamental impact on the optimization process and resulting image fidelity (Spadea et al., 2021).

A broader temporal perspective is provided by Dayarathna et al. (2024), whose survey spans deep learning-based medical image synthesis from 2018 to 2023 (note that it not only covers sCT, but also synthetic MRI and synthetic positron emission tomography (PET) tasks). As illustrated in Figure 2.3, CNN-based models initially dominated, followed by a marked increase in GAN adoption, and more recently a shift toward Transformer and Diffusion-based architectures. Among generative models, cGANs and CycleGANs have been the most widely utilized for sCT generation.

Although CNN-based generative models have achieved considerable success in cross-modality image synthesis, their limited receptive field restricts them to local feature extraction, making it difficult to capture long-range anatomical dependencies. To address this, Transformer and

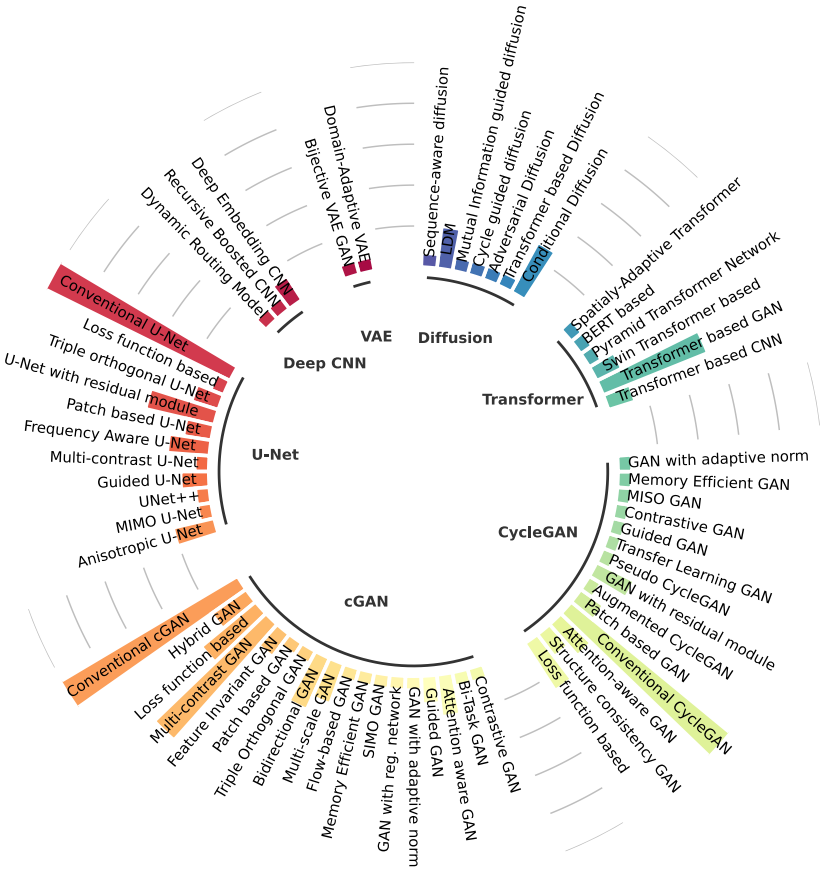


Figure 2.3: TAXONOMY OF DEEP LEARNING ARCHITECTURES FOR MEDICAL IMAGE SYNTHESIS. An overview of deep neural network architectures for medical image synthesis in the literature (Dayarathna et al., 2024).

attention-based architectures have gained increasing traction, employing self-attention mechanisms to integrate local and global contextual information across image regions (Dayarathna et al., 2024). Beyond pure self-attention, several works have targeted localized image quality through mechanisms such as dual-stream generators and Grad-CAM guidance, helping networks focus on anatomically relevant regions for sharper textures and better structural consistency. Hybrid architectures combining CNNs and Transformers have proven especially effective in balancing local detail preservation and global context: Zhao et al. (2023) implemented a Transformer-based bottleneck within an encoder-decoder generator, stabilizing training and enhancing fine structural representation, while ResViT (Dalmaz et al., 2022) employed stacked aggregated residual Transformer layers with weight sharing to reduce computational load. Li et al. (2023c) proposed a multi-scale CNN–Transformer hybrid extracting low-level spatial information through convolutional encoders and modeling global dependencies in a Transformer bottleneck. Zeng et al. (2022) replaced linear projections with convolutional layers to preserve spatial locality, and Li et al. (2023b) combined cross-modality attention with CNN–Transformer integration for improved contextual correlation modeling across MRI contrasts. Yan et al. (2022) utilized multiple residual Swin Transformer blocks interleaved with convolutional layers to enhance local feature learning.

Diffusion-based networks have more recently emerged as one of the most powerful generative frameworks in medical imaging, conditioning the reverse denoising process on source modality data via either classifier-guided (Dhariwal and Nichol, 2021) or classifier-free (Ho and Salimans, 2022) strategies-with most medical synthesis works adopting the latter. Lyu and Wang (2022) demonstrated enhanced realism over GAN baselines using an MRI-conditioned diffusion model, while Li et al. (2023a) incorporated null-space CT components into the reverse process for more precise structural recovery. Meng et al. (2022) introduced multi-modality conditioning to handle missing input modalities, and Jiang et al. (2023) proposed CoLa-Diff, a structure-guided model using anatomical masks with adaptive conditioning weighting. To improve computational efficiency, several works adopted Latent Diffusion Models (LDMs), performing diffusion in compressed latent spaces (Dayarathna et al., 2024). Özbey et al. (2023) introduced an adversarial diffusion model trained in an unsupervised manner, and Zhu et al. (2023) extended 2D diffusion to volumetric synthesis via slice mapping with 3D fine-tuning. The combination of diffusion and Transformer architectures has proven particularly effective: Pan et al. (2024) integrated a Swin Transformer encoder–decoder into a 3D diffusion model conditioned on MRI volumes, significantly outperforming GAN-based baselines by capturing fine structural details through sequential denoising. Together, Transformer and Diffusion-based architectures mark a new generation of sCT synthesis models capable of combining spatial awareness, high structural fidelity, and global anatomical consistency-key aspects for reliable clinical adoption (Dayarathna et al., 2024).

2.3.2 Literature Review: Challenges

Although deep learning has substantially improved the ability to learn complex nonlinear relationships between imaging modalities, several challenges persist across different network designs.

Architectural Limitations Traditional CNN-based architectures for medical image synthesis can struggle to capture fine-grained contextual details due to their fixed receptive fields, particularly when operating on small input sizes. While GANs have achieved notable success, they also face challenges such as training instability, mode collapse, and limited ability to model highly complex data distributions. Vision Transformers (ViTs), though effective at capturing global dependencies, may fail to extract multi-scale local details when operating on fixed-size patches. Diffusion models, despite their impressive image fidelity, can sometimes alter the original data distribution through noise injection, risking the loss of structural consistency (Dayarathna et al., 2024).

2D vs. 3D Processing Many network architectures process 2D slices extracted from 3D volumes, which can lead to discontinuity between slices and the loss of voxel-wise spatial correlations. Conversely, 3D architectures preserve spatial consistency but require substantially more training data, computational resources, and trainable parameters. In patch-based learning, the choice of patch size and overlap ratio is equally critical: too small a patch can disrupt contextual relationships between organs and tissues, while larger patches dramatically increase memory and computational costs (Dayarathna et al., 2024).

Local vs. Global Feature Modeling Cross-modality translation often struggles to balance local texture accuracy and global structural coherence. Purely CNN-based models are typically strong at local feature extraction but weak at modeling long-range dependencies. Conversely, Transformers excel at capturing global context but may overlook fine spatial details. To reconcile these complementary strengths, hybrid CNN-Transformer designs have been widely adopted (Dayarathna et al., 2024).

Paired Data and Misalignment Supervised image-to-image translation methods require paired and spatially aligned datasets, which are often difficult to obtain in medical imaging. Misalignments between MRI and CT pairs can lead to unrealistic translations, geometric distortions, or shifted anatomical boundaries. While CycleGANs mitigate the need for paired data by enabling unpaired training, maintaining precise structural consistency remains an open challenge (Dayarathna et al., 2024).

Data Heterogeneity A significant challenge that cuts across all architectures and most past evaluations concerns generalization: most models are trained on datasets from single centers, limiting their transferability across institutions or imaging protocols and restricting the development of universal models for medical imaging (Dayarathna et al., 2024).

Computational Cost Most deep architectures for medical image synthesis are computationally demanding, especially when dealing with high-dimensional volumetric data. Limited GPU resources can lead to long training times or force models to rely on lower-dimensional inputs, resulting in degraded spatial consistency. Recent architectures such as Transformers and Diffusion models provide superior results but at the cost of significantly increased computational overhead (Dayarathna et al., 2024). Common strategies to mitigate these costs include employing patch-based learning and neighboring-slice aggregation to simulate 3D context efficiently (Zhao et al., 2021; Oulbacha and Kadoury, 2020), using weight-sharing mechanisms and lightweight modules to reduce trainable parameters (Kläser et al., 2018; Dalmaz et al., 2022), extending 2D networks to pseudo-3D through stacked slice or volumetric layer integration (Zhu et al., 2023), training models in latent space via LDMs to decrease dimensionality while maintaining fidelity (Jang et al., 2023; Jiang et al., 2023; Zhu et al., 2023), and combining Transformers with CNN bottlenecks to leverage global attention with reduced computational intensity (Dalmaz et al., 2022). Hierarchical designs such as Swin Transformers further improve memory efficiency over vanilla ViTs (Yan et al., 2022).

2.4 Fullbody Approaches

Many existing approaches rely on region-specific models, but there are reasons for developing generalized or *fullbody* models capable of synthesizing CT images across multiple anatomical regions, including the thorax, abdomen, and pelvis. A single model trained on diverse anatomical regions could simplify deployment, eliminate the need for multiple site-specific networks, and enhance consistency across treatment areas. However, achieving this is challenging due to anatomical variability, motion artifacts, and scanner-dependent differences.

To date, only a few studies have explicitly addressed region-independent or fullbody sCT generation. Among these, Hsu et al. (2025) conducted one of the first feasibility studies using a general cGAN model trained jointly on thoracic, abdominal, and pelvic MRI data, achieving image and dosimetric accuracy comparable to that of site-specific models. Their findings indicate that a single deep learning model could potentially enable MRI-only workflows across multiple extracranial anatomical regions. Similarly, one of the top-performing SynthRad2025 teams (Mei et al., 2025) trained a single model on the combined dataset rather than developing separate site-specific models for each anatomical region.

While explicit comparisons between site-specific and general models remain limited in the MRI-sCT literature, there is a clear trend toward improving model generalizability across imaging systems and acquisition domains. Rather than developing separate networks for each setting, recent studies have focused on making models more robust to scanner-, sequence-, and field-strength variability. For example, Fetty et al. (2020) demonstrated that a cGAN trained on 0.35 T

MRI could be successfully transferred to 1.5 T and 3 T systems with minimal performance loss, highlighting the feasibility of cross-field generalization. Similarly, Brou Boni et al. (2020) achieved consistent pelvic sCT generation across multiple institutions and scanners.

Despite the growing interest in generalizable or fullbody sCT models, several expert reviews highlight that a site-specific approach remains preferable for clinical accuracy and safety. The recent position paper by the European Society for Radiotherapy and Oncology (ESTRO; Villegas et al. 2024) emphasizes that the trade-off between accuracy and generalizability should be handled with caution, explicitly noting that the preferable approach is site-specific, as each anatomical region presents unique imaging characteristics, motion patterns, and dose calculation requirements. The authors argue that training data should be carefully curated to represent the intended clinical cohort – for example, "male pelvis without hip implants" – and that models should not be applied outside their validated anatomical or scanner context. They further highlight that scanner-dependent differences, field-of-view limitations, and bone-air interface artifacts can severely compromise general model performance. Consequently, they advocate for separate model development and commissioning per anatomical site, supported by rigorous site-specific quality assurance and risk analysis frameworks. Investigating general models still has scientific value. Comparing multi-region training strategies against site-specific baselines is the only way to measure the accuracy trade-off and determine when general models are clinically acceptable.

2.5 SynthRAD2023 and 2025 Challenge

To improve and benchmark the creation of sCT images using MRI and cone beam computed tomography (CBCT) data for radiation, the SynthRAD Grand Challenges 2023[1] and 2025[2] were created. A basic gap in adaptive radiotherapy workflows was filled by the challenge: CBCT helps determine the patient's position on the treatment table but needs precise attenuation data, while MRI provides good soft-tissue contrast but lacks the electron density information necessary for dose calculations. The goal of combining CT from multiple modalities is closing that gap and possibly lessen dependency on traditional CT scans.

Two main tasks were defined:

- **Task 1:** Generate sCT from MRI (MRI-to-CT synthesis), supporting MRI-only or MR-guided radiotherapy applications.
- **Task 2:** Generate sCT from CBCT (CBCT-to-CT synthesis), enabling CBCT-based adaptive radiotherapy workflows.

While the tasks of both challenge editions are very similar, two key differences exist between the 2023 and 2025 editions. The 2023 edition prohibited the use of external training data and pretrained models, whereas the 2025 edition permitted these provided they were publicly available to all participants before a specified date. Additionally, the provided datasets cover different anatomical regions: brain and pelvis in 2023, and head-and-neck, thorax, and abdomen in 2025.

The following summarizes the methodological approaches of the top five SynthRAD2025 submissions for Task 1 (MRI-to-sCT), covering architectural choices, loss function design, and use of pretrained models.

The winning group *KoalAI* (Xin et al., 2025) employed an ensembled 3D ResUNet-L with a Masked Anatomical Perception (MAP) loss, which combines MSE with a perceptual term from a TotalSegmentator model to enforce anatomical consistency between synthetic and real CT.

[1] https://synthrad2023.grand-challenge.org/
[2] https://synthrad2025.grand-challenge.org/

The second-placed group *ImagePasNet* (González et al., 2025) adapted a 3D residual UNet from nnUNet and introduced the Anatomical Feature-Prioritized (AFP) loss, which supervises training using multi-layer features from a TotalSegmentator-based segmentation network.

The third-placed group *BreizhCT* (Boussot et al., 2025) used a 2.5D UNet++ with a ResNet-34 encoder, trained jointly across regions and fine-tuned per region, with a loss combining L_1 error and perceptual terms from VGG-19, SAM, and TotalSegmentator.

The fourth-placed group *MixCT* (Mei et al., 2025) proposed GANeXt, a 3D patch-based GAN built from ConvNeXt blocks, notable as the only top submission to train a single unified model across all anatomical regions without region-specific fine-tuning.

The fifth-placed group *QWER* (Han and Tan, 2025) applied the 3D nnUNet framework, combining MAE with a TotalSegmentator-based perceptual loss, and achieved their best results by ensembling region-specific models with test-time augmentation (TTA).

Across the top five submissions, several consistent patterns emerge. The three best-performing teams all employed encoder-decoder architectures based on UNet variants, confirming the continued effectiveness of this design for cross-modality image synthesis. Their primary differences arose not from architectural complexity but from the choice and composition of loss functions: the highest-performing models incorporated deep feature representations from pretrained networks – most notably TotalSegmentator (discussed in more detail in Section 3.2) – as perceptual and anatomical supervision signals, improving training stability and geometric consistency. Notably, all top methods except MixCT trained distinct models per anatomical region, suggesting that current approaches still rely on region-specific optimization rather than fullbody generalization – a limitation that directly motivates the multi-region scope of this work.

Chapter 3

Background

This chapter introduces the technical foundations underlying the methods developed in this report. It begins with an overview of MRI and CT imaging principles and the motivation for MRI intensity normalization, before covering the image-to-image translation frameworks and generator network architectures that form the basis of the sCT generation pipeline.

3.1 MR and CT Imaging

MRI is a non-ionizing imaging modality that provides high-resolution, three-dimensional visualization of soft tissues. Image formation is based on the interaction of hydrogen protons in water molecules with a strong external magnetic field and radiofrequency pulses. As the protons relax back to equilibrium, they emit signals that are detected and reconstructed into images reflecting tissue-specific properties such as water and fat content (Smith and Webb, 2010). This results in excellent soft-tissue contrast, enabling clear differentiation between tumors and surrounding organs at risk, which is often limited in CT images. For certain clinical indications, contrast agents may be used to further enhance tissue contrast and vascular structures. Because MRI uses radiation in the radiofrequency range, which does not damage tissue as it passes through, there are no known biological hazards associated with the modality (AlHadidi et al., 2026).

CT imaging, in contrast, relies on ionizing X-ray radiation. The fundamental principle behind image generation is the differential absorption and attenuation of X-rays by tissues of varying density. Dense structures such as bone appear with high attenuation, while soft tissue and air-filled regions show lower attenuation. Because CT uses ionizing radiation, it is associated with exposure that carries a potential biological risk, particularly with repeated imaging (AlHadidi et al., 2026). A key advantage of CT in radiotherapy is its ability to provide quantitative electron density information, which is essential for accurate dose calculation. Treatment planning systems use CT-derived HU (Hounsfield, 1980) to assign electron densities relative to water, enabling reliable modeling of radiation transport and energy deposition. For this reason, CT has long been considered the gold standard for radiotherapy treatment planning (Smith and Webb, 2010).

These HU form a standardized scale derived from tissue-specific X-ray attenuation:

$$Y'(v) = 1000 \cdot \frac{\mu(v) - \mu_{\mathrm{H_2O}}}{\mu_{\mathrm{H_2O}}}, \tag{3.1}$$

where $Y'(v)$ is the CT number in HU at voxel v, $\mu(v)$ the linear attenuation coefficient of the tissue at voxel v, and $\mu_{\mathrm{H_2O}}$ that of water. By construction, water is assigned a value of 0 HU and air approximately -1000 HU. Because the scale is anchored to these physical reference points, CT intensities are quantitative and largely comparable across scanners following calibration. Table 3.1

Table 3.1: This table lists representative CT numbers for common tissue types at a typical clinical energy range (effective energy ≈60-70 keV), adapted from Smith and Webb (2010).

Tissue	CT number (HU)
Bone	1000 to 3000
Blood	40
Brain (grey matter)	35 to 45
Brain (white matter)	20 to 30
Muscle	10 to 40
Water	0
Lipid	-50 to -100
Air	-1000

lists representative CT numbers for common tissue types at a typical clinical energy range (Smith and Webb, 2010).

Conventional MRI, by contrast, lacks an absolute intensity scale. MR signal intensities are unitless quantities expressed in uncalibrated units – they do not correspond to a defined physical measure and their magnitude depends on hardware-specific scaling factors such as receiver gain and coil sensitivity (Smith and Webb, 2010). The observed intensity of a given voxel arises from a complex interaction between tissue properties (proton density, T_1, T_2) and acquisition parameters (field strength, pulse sequence, repetition and echo times). Consequently, the same tissue can yield substantially different numerical values across scanners, field strengths, or even repeated acquisitions on the same device with a different protocol; MRI intensities are therefore not directly comparable across – or even within – datasets without explicit normalization. The use of contrast agents, which alter tissue relaxation times, introduces an additional source of variability. This fundamental absence of a standardized intensity scale motivates the application of normalization techniques when MRI data serve as input for models that aim to recover physically meaningful quantities, such as sCT images.

3.1.1 Nyúl Histogram Matching Normalization

Nyúl histogram standardization is a widely used method for MRI intensity normalization that aims to reduce inter-subject and inter-scanner variability by mapping volume intensities to a common standard scale. The method was originally proposed by Nyúl and Udupa (1999) and is based on the observation that although absolute MRI intensities are not standardized, the relative positions of tissue classes within the intensity histogram are often consistent across subjects. The core idea is to learn a global intensity mapping from a representative training set and subsequently apply this mapping to unseen volumes. Unlike per-volume normalization methods, which operate independently on each volume, Nyúl normalization establishes a dataset-level reference scale, enabling consistent intensity interpretation across the entire cohort. The method proceeds in two stages: a training stage, in which a standard histogram is learned, and a transformation stage, in which individual images are mapped to this standard.

Training stage. Given a set of N training volumes, the algorithm first identifies, for each volume v, a set of landmark intensities derived from the foreground intensity histogram. Specifically, let $p_{\min}$ and $p_{\max}$ denote user-defined percentiles that delineate the effective intensity range, excluding background voxels and extreme outliers. Within the interval $[p_{\min}, p_{\max}]$, a fixed number L of equally spaced percentile levels are computed, and the corresponding intensity values at those percentile levels are recorded as landmarks $\mu_1^{(v)}, \ldots, \mu_L^{(v)}$ for each volume v, where $\mu_l^{(v)}$ de-

notes the intensity at the l-th percentile landmark of volume v. In a second step, the landmark values are averaged across all training volumes to obtain a standard set of landmark intensities:

$$\bar{\mu}_l = \frac{1}{N} \sum_{v=1}^{N} \mu_l^{(v)}, \quad l = 1, \ldots, L. \tag{3.2}$$

Together with the desired output minimum s_{min} and output maximum s_{max}, these averaged landmarks define the standard histogram scale.

Transformation stage. To normalize a volume, the same percentile landmarks $\mu_1, \ldots, \mu_L$ are extracted from its foreground histogram. A piecewise linear mapping is then constructed that maps each volume-specific landmark μ_l to the corresponding standard landmark $\bar{\mu}_l$, with linear interpolation between consecutive landmarks. Intensities below $p_{\min}$ are clipped to $s_{\min}$, and intensities above $p_{\max}$ are clipped to $s_{\max}$. Formally, for a voxel at position $\mathbf{v}$ of the MRI before normalization with intensity $X'(\mathbf{v})$ falling in the interval $[\mu_l, \mu_{l+1}]$, the normalized intensity is:

$$X(\mathbf{v}) = \bar{\mu}_l + \frac{X'(\mathbf{v}) - \mu_l}{\mu_{l+1} - \mu_l} \left(\bar{\mu}_{l+1} - \bar{\mu}_l\right). \tag{3.3}$$

A binary foreground mask is used throughout to restrict the histogram computation to anatomically relevant voxels and to exclude background air, which would otherwise dominate the intensity distribution and distort the landmark estimates.

3.1.2 N-Peaks Intensity Normalization

N-Peaks is a tissue-based MRI intensity normalization method proposed by Wallimann et al. (2025) that harmonizes image intensities across a dataset by aligning the histogram peak intensities of selected homogeneous normal tissues. Unlike the previously described Nyúl histogram matching, which operates on aggregate intensity statistics of the full body contour, N-Peaks anchors the transformation to anatomically meaningful tissue landmarks. This makes it less susceptible to bias from inter-patient variability in body composition, such as differences in fat volume, which can distort global histogram-based normalizations (Wallimann et al., 2025).

The method takes as input the non-normalized MR image X', and a set of K tissue masks $M_j \subset X'$, $j = 1, \ldots, K$, each expected to contain a region of homogeneous normal tissue. In a first step, the algorithm isolates homogeneous voxels within each mask by computing a local intensity change (LIC) map. For each voxel v with intensity $X'(v)$, this is defined as:

$$g_r(v) \;=\; \max_{w \,\in\, W_r(v)} X'(w) \;-\; \min_{w \,\in\, W_r(v)} X'(w)\,, \qquad \text{where} \quad W_r(v) = \left\{\, v' \in X' \;\middle|\; d(v,\, v') \le r \,\right\}, \tag{3.4}$$

and d denotes the Euclidean distance with r being a user-defined neighborhood radius. For each mask M_j, only voxels with local intensity change below a threshold g_j are retained to form the homogeneous subregion (subscript h for homogeneous):

$$M_{h,j} \;=\; \left\{\, v \in M_j \;\middle|\; g_r(v) \le g_j \,\right\}. \tag{3.5}$$

The threshold g_j is set to the q-th quantile of the LIC values within M_j, where $q \in (0, 1]$ controls the strictness of filtering. A lower q yields a stricter g_j, retaining only the most homogeneous voxels at the cost of reduced sample size, while a higher q applies softer filtering and retains more voxels. This filtering excludes tissue interfaces, blood vessels, and other heterogeneous structures.

Next, for each filtered region $M_{h,j}$, a landmark intensity μ_j is identified by computing a kernel density estimate (KDE) of the intensity distribution within $M_{h,j}$ and selecting the intensity at the most prominent peak:

$$\mu_j \;=\; \arg\max_{x} \; \hat{f}_{h,j}(x)\,, \tag{3.6}$$

where $\hat{f}_{h,j}$ denotes the KDE computed over the voxel intensities in $M_{h,j}$, and a prominence criterion is applied to handle potentially multi-modal distributions. This peak detection is robust against imperfect masks, as the algorithm can isolate the correct tissue peak even when the mask contains multiple tissue types (Wallimann et al., 2025).

Once landmark intensities μ_j have been determined for all K masks, the image is transformed to a normalized scale via a piecewise linear mapping. Each landmark is mapped to a predefined target intensity $\bar{\mu}_j$, analogous to the standard landmarks in Nyúl normalization (Equation 3.3). For an intensity value $X'(v)$ falling in the interval $[\mu_j,\, \mu_{j+1}]$, the normalized intensity is:

$$X(v) \;=\; \bar{\mu}_j + \frac{X'(v) - \mu_j}{\mu_{j+1} - \mu_j} \left(\bar{\mu}_{j+1} - \bar{\mu}_j\right). \tag{3.7}$$

The transformation is thus identical in form to the Nyúl piecewise linear mapping (Equation 3.3), but differs in how landmarks are determined: Nyúl uses histogram percentiles of the body contour, whereas N-Peaks uses tissue-specific peak intensities from anatomically defined regions.

3.1.3 N4 Bias Field Correction

MRI images are affected by a slowly varying multiplicative signal inhomogeneity known as the bias field (also referred to as intensity inhomogeneity or shading artifact), which arises primarily from non-uniformities in the transmit and receive radiofrequency coils. This artifact causes the intensity of the same tissue type to vary smoothly across the image volume, violating the assumption of spatially uniform tissue intensities that underlies histogram-based normalization methods. If left uncorrected, the bias field broadens and distorts tissue peaks in the intensity histogram, reducing the ability of methods such as N-Peaks to identify sharp, well-separated tissue landmarks.

This can be mitigated by applying N4ITK bias field correction (Tustison et al., 2010) to MRI volumes prior to intensity normalization. N4ITK is a widely adopted refinement of the earlier N3 algorithm. It assumes that the observed voxel intensity $i(v)$ is a corrupted version of the true, homogeneous tissue intensity $X'_{\text{true}}(v)$:

$$X'(v) \;=\; X'_{\text{true}}(v) \cdot b(v) + n(v)\,, \tag{3.8}$$

where $b(v)$ is a smooth, slowly varying multiplicative bias field and $n(v)$ denotes additive noise. The algorithm iteratively estimates and removes $b(v)$ by fitting a B-spline approximation to the log-transformed residual field, using a modified expectation-maximization scheme that does not require prior tissue class information.

3.2 TotalSegmentator

TotalSegmentator (Wasserthal et al., 2023; Akinci D'Antonoli et al., 2025) is a publicly available framework for whole-body segmentation in medical imaging. It is built on the nnUNet architecture (Isensee et al., 2021), which self-configures preprocessing, network topology, and training hyperparameters based on the properties of the input data. The framework automatically delineates over one hundred anatomical structures (so-called *tasks*) from volumetric CT and MRI scans and accepts 3D input volumes at varying spatial resolutions; internally, images are resampled to a target spacing determined during the self-configuration stage. For MRI, a dedicated model trained on MR data is available, supporting tissue segmentation tasks such as organ and body-composition delineation.

3.3 GAN-based Image-to-Image Translation

Image-to-image translation refers to the task of learning a mapping between two visual domains, where both input and output are images defined on a grid. Formally, given paired 2D samples (x, y) from domains $\mathcal{X}$ and $\mathcal{Y}$, the goal is to learn a function $G : \mathcal{X} \rightarrow \mathcal{Y}$ such that $G(x)$ approximates the corresponding target image y. In the context of this project, $\mathcal{X}$ corresponds to the domain of normalized MR images and $\mathcal{Y}$ to domain of normalized CT images.

3.3.1 Pix2Pix (cGAN)

Isola et al. (2017) proposed Pix2Pix which formulates image-to-image translation as a *conditional generative adversarial network* (cGAN). The key difference compared to conventional GANs is, that instead of generating images from random noise alone, the generator is conditioned on an input image x and learns a mapping $G : \{x, z\} \rightarrow y$, where z denotes optional noise. In practice, stochasticity is often introduced via dropout rather than explicit noise injection (Isola et al., 2017).

Network Components. Pix2Pix consists of two networks, schematically depicted in Figure 3.1. The **Generator** G translates normalized MR to the normalized CT domain using an CED architecture with skip connections (UNet, see Section 3.4.1). Skip connections concatenate feature maps between corresponding encoder and decoder layers, allowing low-level spatial information to bypass the bottleneck (Isola et al., 2017). This is particularly beneficial for tasks where input and output share structural alignment, such as MR-to-CT translation. The **Discriminator** D_Y classifies each $P \times P$ patch as real or generator-generated and averages the responses. Pix2Pix uses a *PatchGAN* classifier that operates on local image patches instead of the full image. This design models high-frequency structure while reducing the number of parameters and enabling fully convolutional inference (Isola et al., 2017).

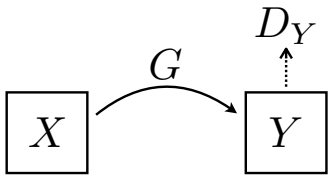


Figure 3.1: PIX2PIX NETWORK ARCHITECTURE. Schematic visualization of the Pix2Pix network architecture, adapted from Zhu et al. (2017).

Objective. The conditional adversarial objective is based on two components, the *cGAN* and L_1 objectives, and is defined as:

$$G^* = \arg\min_G \max_{D_Y} \mathcal{L}_{\text{cGAN}}(G, D_Y) + \lambda \mathcal{L}_{L_1}(G), \tag{3.9}$$

where λ controls the trade-off between structural realism and pixel-wise fidelity (Isola et al., 2017).

The cGAN objective is defined as

$$\mathcal{L}_{\text{cGAN}}(G, D_Y) = \mathbb{E}_{x,y}[\log D_Y(x, y)] + \mathbb{E}_{x,z}[\log(1 - D_Y(x, G(x, z)))], \tag{3.10}$$

where the discriminator D_Y learns to distinguish real pairs (x, y) from generator-generated pairs $(x, G(x, z))$ in domain $\mathcal{Y}$, and the generator G aims to minimize this objective while D_Y maximizes it. Conditioning the discriminator on x is essential, as it enforces consistency between input and output rather than merely encouraging realism in $\mathcal{Y}$ (Isola et al., 2017).

To stabilize training and preserve low-frequency correctness, Pix2Pix combines the adversarial loss with a pixel-wise reconstruction term. Instead of an L_2 loss, an L_1 loss is used to reduce blurring:

$$\mathcal{L}_{L_1}(G) = \mathbb{E}_{x,y,z}\big[\|y - G(x, z)\|_1\big]. \tag{3.11}$$

Data Requirements. Pix2Pix is a *supervised* framework and requires paired and spatially aligned training data (x, y). In our application, this corresponds to registered MR–CT pairs. The availability and quality of such paired datasets directly influence the achievable performance, which is a major limitation of Pix2Pix (Boulanger et al., 2021).

3.3.2 CycleGAN

CycleGAN extends adversarial image-to-image translation to the *unpaired* setting, where no spatially aligned training pairs (x, y) are available (Zhu et al., 2017). Instead, only two independent image collections $x_i \in \mathcal{X}$ and $y_j \in \mathcal{Y}$ are given. The goal is to learn mappings between the two domains solely from distribution-level supervision.

Network Components. CycleGAN consists of four networks (see Fig. 3.2): Two generators – $G : \mathcal{X} \to \mathcal{Y}$ and $F : \mathcal{Y} \to \mathcal{X}$ – and two discriminators – D_Y distinguishing real images y from generated images $G(x)$ and D_X distinguishing real images x from generated images $F(y)$ (Zhu et al., 2017).

Zhu et al. (2017) implemented both generators as fully convolutional encoder-decoder networks with residual blocks (see Section 3.4.2), while the discriminators followed the PatchGAN design introduced in Pix2Pix (see Section 3.3.1). The architecture enables translation at the level of local image patches while preserving global consistency.

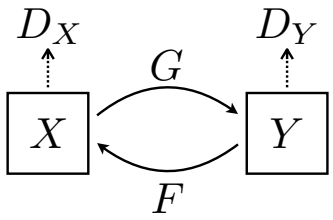


Figure 3.2: CycleGAN Network Architecture. Schematic visualization of the CycleGAN network architecture from Zhu et al. (2017).

Objective. The optimization problem combines adversarial and cycle-consistency terms:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \mathcal{L}_{\text{GAN}}(G, D_Y, \mathcal{X}, \mathcal{Y}) + \mathcal{L}_{\text{GAN}}(F, D_X, \mathcal{Y}, \mathcal{X}) + \lambda \mathcal{L}_{\text{cyc}}(G, F), \tag{3.12}$$

where λ controls the trade-off between distribution matching and reconstruction consistency.

The adversarial loss for the mapping $G : \mathcal{X} \to \mathcal{Y}$ is defined as

$$\mathcal{L}_{\text{GAN}}(G, D_Y, \mathcal{X}, \mathcal{Y}) = \mathbb{E}_{y\sim\mathcal{Y}}[\log D_Y(y)] + \mathbb{E}_{x\sim\mathcal{X}}[\log(1 - D_Y(G(x)))], \tag{3.13}$$

where G aims to generate images indistinguishable from real samples in $\mathcal{Y}$, while D_Y maximizes its discrimination accuracy. Note the difference to the cGAN objective (see Equation (3.10)), where the discriminator is conditioned on the input. An analogous objective $\mathcal{L}_{\text{GAN}}(F, D_X, \mathcal{Y}, \mathcal{X})$ is defined for the reverse direction (Zhu et al., 2017).

Adversarial training alone enforces matching marginal distributions but does not guarantee that an individual input x is mapped to a semantically corresponding output. To constrain the solution space, CycleGAN introduces a *cycle-consistency* constraint (Zhu et al., 2017). The intuition is that a translation from one domain to the other and back again should reconstruct the original image $x \to G(x) \to F(G(x)) \approx x$ and $y \to F(y) \to G(F(y)) \approx y$. This is formalized as

$$\mathcal{L}_{\text{cyc}}(G, F) = \mathbb{E}_{x\sim\mathcal{X}}\big[\|F(G(x)) - x\|_1\big] + \mathbb{E}_{y\sim\mathcal{Y}}\big[\|G(F(y)) - y\|_1\big]. \tag{3.14}$$

Data Requirements. In contrast to Pix2Pix, CycleGAN does not require paired or spatially aligned training data. Instead, it only assumes that samples from both domains are available and that an underlying correspondence exists at the distribution level (Zhu et al., 2017). This property makes CycleGAN particularly attractive for medical imaging applications where perfectly registered MR–CT pairs are scarce or costly to obtain. Interestingly, Wolterink et al. (2017) have shown that models trained on unpaired data can even outperform paired-data GANs in certain scenarios.

3.3.3 Contrastive Unpaired Translation (CUT)

CUT, proposed by Park et al. (2020), reformulates unpaired image-to-image translation by replacing the cycle-consistency constraint of CycleGAN with a contrastive learning objective. While CycleGAN enforces structural preservation through a full round-trip reconstruction involving two generators and two discriminators, CUT achieves this with only a single generator $G : \mathcal{X} \to \mathcal{Y}$ and a single discriminator D_Y.

The key insight behind CUT is that content preservation between input and output can be enforced at the level of internal feature representations rather than through pixel-wise reconstruction in image space. Specifically, CUT leverages the observation that corresponding spatial locations in the input image x and the generated output $G(x)$ should share similar high-level features, while non-corresponding locations should differ. This is formalized through a patch-based contrastive loss applied to intermediate feature maps of the generator itself (Park et al., 2020).

Network Components. CUT consists of the following three components. The **Generator** G is a CED network with residual blocks. In CUT, the generator additionally serves as a feature extractor for the contrastive loss: intermediate encoder layers produce the feature maps from which patch-level queries and keys are derived (Park et al., 2020). The **Discriminator** D_Y serves the same purpose and utilizes the same PatchGAN architecture as in Pix2Pix and CycleGAN. The **Feature projection head** H is an additional, small two-layer MLP network where H_l is attached to each selected encoder layer l to project the extracted features into a lower-dimensional embedding space where the contrastive loss is computed. These projection heads are learned jointly with the generator (Park et al., 2020).

Objective. The full CUT objective combines an adversarial loss with the PatchNCE contrastive loss:

$$G^* = \arg\min_G \max_{D_Y} \mathcal{L}_{\text{GAN}}(G, D_Y, \mathcal{X}, \mathcal{Y}) + \lambda_{NCE}\mathcal{L}_{\text{PatchNCE}}(G, H, \mathcal{X}), + \lambda_{NCE}\mathcal{L}_{\text{PatchNCE}}(G, H, \mathcal{Y}), \tag{3.15}$$

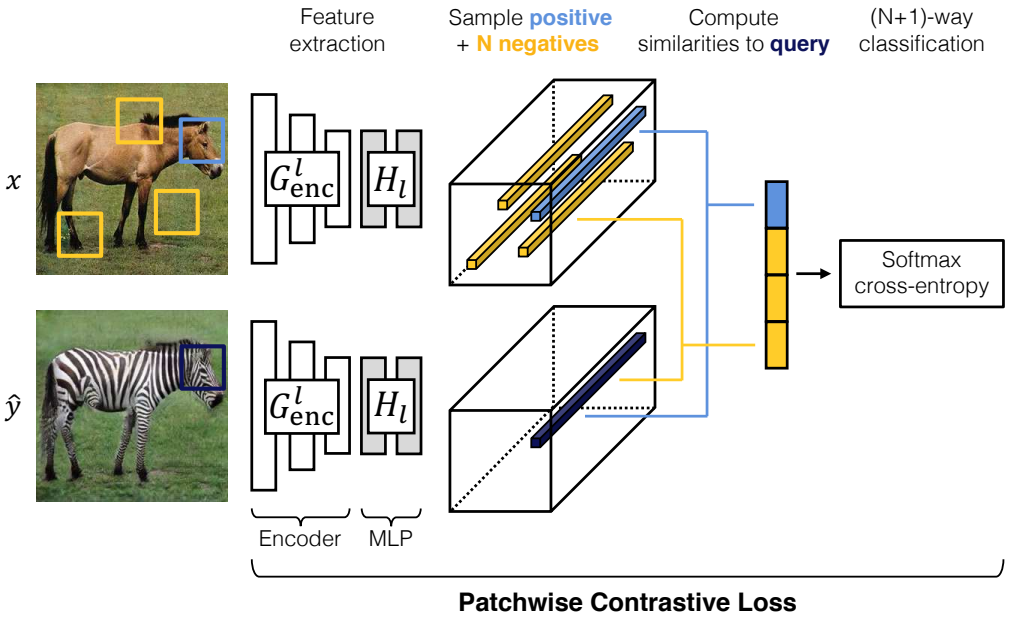


Figure 3.3: CONTRASTIVE UNPAIRED TRANSLATION: PATCHWISE CONTRASTIVE LOSS. For each query patch in the generated output, the corresponding patch in the input image x serves as the positive sample, while all other spatial locations within x serve as N negatives. Features are extracted via a shared encoder G^l_{enc} and projected through an MLP H_l (Park et al., 2020).

where λ_{NCE} controls the relative weight of the contrastive term (Park et al., 2020). The adversarial loss $\mathcal{L}_{\text{GAN}}$ follows the same formulation as in CycleGAN (see Equation (3.13)).

The PatchNCE loss is the central contribution of CUT. For a selected encoder layer l of the generator, let $\mathbf{z}_l^{(s)} = H_l(G^l_{\text{enc}}(G(x)))^{(s)}$ denote the projected feature at a spatial location s in the generated output $G(x)$, and let $\mathbf{z}_l^{+(s)} = H_l(G^l_{\text{enc}}(x))^{(s)}$ denote the projected feature at the same spatial location s in the input x. These form a positive pair: corresponding patches in input and output that should be mapped to similar representations. All features at other spatial locations $s \neq s'$ in the input serve as negative samples. The PatchNCE loss for a single query at layer l and location s is defined as:

$$\ell(s,l) = -\log \frac{\exp(\mathbf{z}_l^{(s)} \cdot \mathbf{z}_l^{+(s)}/\tau)}{\exp(\mathbf{z}_l^{(s)} \cdot \mathbf{z}_l^{+(s)}/\tau) + \sum_{s' \neq s} \exp(\mathbf{z}_l^{(s)} \cdot \mathbf{z}_l^{-(s')}/\tau)}, \tag{3.16}$$

where τ is a temperature parameter and $\mathbf{z}_l^{-(s')}$ are the negative samples drawn from other spatial locations within the same input image. A illustration of the patchwise contrastive loss can be seen in Figure 3.3. The total PatchNCE loss is computed by averaging over S randomly sampled spatial locations and L selected encoder layers (Park et al., 2020):

$$\mathcal{L}_{\text{PatchNCE}}(G, H, \mathcal{X}) = \mathbb{E}_{x \sim \mathcal{X}} \left[\frac{1}{L} \sum_{l=1}^{L} \frac{1}{S} \sum_{s=1}^{S} \ell(s,l) \right]. \tag{3.17}$$

A notable property of this formulation is that both positive and negative samples are drawn from within the same image. Since the generator encoder is used both to produce the output and

to extract features for the contrastive loss, the framework effectively encourages the generator to preserve the spatial structure encoded in its own intermediate representations (Park et al., 2020).

Data Requirements. Like CycleGAN, CUT is an unsupervised framework that does not require paired or spatially aligned training data. It requires only two independent collections of images from domains $\mathcal{X}$ and $\mathcal{Y}$.

3.4 Generator Network Architectures

The generator constitutes the central component of all image-to-image translation frameworks investigated in this work, including Pix2Pix, CycleGAN and CUT. In this section, we describe the generator architectures used throughout frameworks introduced in Section 3.3.

3.4.1 UNet-256

The UNet architecture was originally proposed for biomedical image segmentation by Ronneberger et al. (2015). It consists of a *contracting path* (encoder) that captures contextual information and a symmetric *expansive path* (decoder) that enables precise spatial localization (see Figure 3.4). A defining architectural element is the use of skip connections that concatenate feature maps from encoder layers with their corresponding decoder layers. This design combines high-resolution spatial information from early layers with semantically rich, low-resolution features propagated through the bottleneck.

While this symmetric encoder-decoder principle with skip connections remains conceptually unchanged, Isola et al. (2017) introduced several architectural modifications to adapt UNet to conditional adversarial image-to-image translation. These adaptations are discussed in the following.

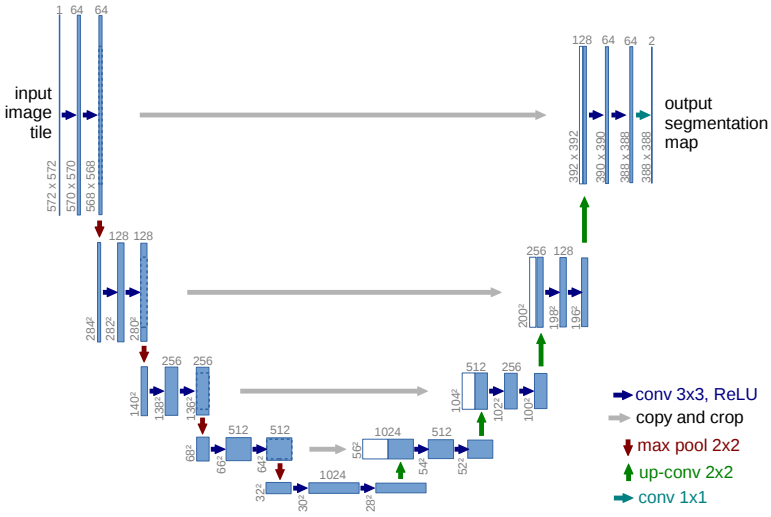


Figure 3.4: ORIGINAL UNET ARCHITECTURE. Original UNet architecture as proposed by Ronneberger et al. (2015). UNet-256 uses 8 levels with 4×4 convolutions with stride 2 to replace max pooling, compressing $256 \times 256 \times 1$ inputs down to a bottleneck of $1 \times 1 \times 512$.

The Pix2Pix framework by Isola et al. (2017) adopted a modified UNet as generator for conditional image-to-image translation. In contrast to the segmentation setting in Ronneberger et al. (2015), the generator learns a mapping between two image domains, and therefore outputs a continuous-valued image rather than discrete class labels.

For 256×256 inputs, Pix2Pix employs an 8-level encoder-decoder, commonly referred to as *UNet-256*. All convolutions use 4×4 kernels with stride 2 for downsampling, effectively replacing max-pooling operations by strided convolutions. Each convolutional block follows a *Convolution* $\rightarrow$ *Normalization* $\rightarrow$ *Activation* pattern. The final layer applies a non-linear activation to constrain the output intensity range.

The encoder of UNet-256 simultaneosly increases the number of channels from 1 (or 3 for natural images) to 512 while reducing the spatial resolution from 256^2 to 1^2. The decoder mirrors this structure using transposed convolutions. Due to the skip connections, feature maps from encoder layer i are concatenated with those of decoder layer $8-i$, doubling the number of channels before the subsequent convolution. Consequently, after concatenation, the number of channels is reduced again by the following transposed convolution. This concatenation mechanism preserves fine-grained spatial information that would otherwise be lost in the bottleneck representation.

While the original UNet used standard convolution and ReLU layers, Pix2Pix introduced normalization layers after each convolution (BatchNorm). Furthermore, dropout is applied in several of the deepest decoder layers. In Pix2Pix, dropout also serves as the only source of stochasticity in the generator and is kept active during inference (Isola et al., 2017).

3.4.2 ResNet-9

The ResNet-9 generator follows the architecture proposed by Johnson et al. (2016) for neural style transfer and subsequently adopted by Zhu et al. (2017) as the default generator for CycleGAN. The ResNet-9 (as shown in Figure 3.5) architecture passes all information through a flat bottleneck composed of nine residual blocks. This design is well suited to tasks where input and output share the same overall structure but differ primarily in texture and appearance, as the residual connections allow the network to learn an additive refinement on top of an approximate identity mapping. The architecture consists of three stages: a downsampling encoder, a residual bottleneck, and an upsampling decoder.

The ResNet-9 encoder begins with a 7×7 convolution preceded by reflection padding, followed by instance normalization and ReLU. Two subsequent 3×3 convolutional layers with stride 2 progressively reduce the spatial resolution by a factor of 4 while doubling the channel count at each stage. Each is followed by instance normalization and ReLU. For 256×256 inputs, the encoder produces feature maps of size $64 \times 64 \times 256$.

The bottleneck consists of nine residual blocks, each operating at 256 channels. Each block contains two 3×3 convolutional layers with reflection padding, where the first is followed by instance normalization and ReLU, and the second by instance normalization only. The output is computed as:

$$\mathbf{h}' = \mathbf{h} + \mathcal{F}(\mathbf{h}), \tag{3.18}$$

where h is the block input and $\mathcal{F}$ the two-layer convolutional transformation. This residual formulation enables the network to learn refinements to its intermediate representations rather than full transformations, facilitating gradient flow and stabilizing training in deeper networks (He et al., 2016). Unlike the skip connections in UNet, which bridge encoder and decoder at matching spatial resolutions, the residual connections here operate within the bottleneck at a single resolution, preserving feature dimensionality without concatenation.

The decoder mirrors the encoder using two upsampling stages that restore the spatial resolution to the original input size. In the anti-aliased variant used here, each stage consists of a learned upsampling operation followed by a 3×3 convolution, normalization, and ReLU activation. The

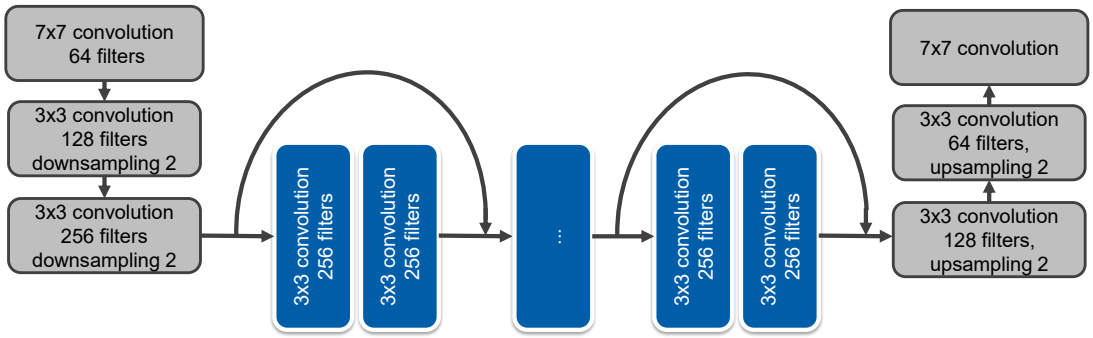


Figure 3.5: RESNET-9 ARCHITECTURE. ResNet-9 architecture as implemented by Zhu et al. (2017). ResNet-9 consists of an encoder, a flat bottleneck with 9 residual blocks, and a decoder restoring original spatial dimensions.

channel dimensions are progressively reduced from 256 to 128 and then to 64. The final output layer applies reflection padding followed by a 7×7 convolution that maps the 64-channel feature representation to the desired number of output channels, and a non-linear output activation constrains the output range.

Park et al. (2020) adopted the ResNet-9 generator as the default for CUT. The sequential structure of ResNet-9 allows straightforward extraction of intermediate feature representations at arbitrary layers, which is required by the PatchNCE loss in CUT (3.17). Park et al. (2020) calculate the PatchNCE loss based on five layers from the encoder and bottleneck. The recursive architecture of UNet-256 does not readily support this.

3.5 Evaluation Metrics

We quantify the similarity between generated sCT volumes and reference CT using masked, slice-wise image-similarity metrics computed inside the patient body region. For each axial slice i in the target volume $y_i \in Y'$ or predicted sCT volume $\hat{y}_i \in \hat{Y}'$, let $y_i(p)$ and $\hat{y}_i(p)$ denote the ground truth CT and sCT intensities (in HU) at pixel p, and let m_i be the corresponding body mask with cardinality $|m_i|$.

Masked mean absolute error (MAE). The masked MAE measures the average absolute deviation between CT and sCT within the body mask:

$$\mathrm{MAE}_i = \frac{1}{|m_i|} \sum_{p \in m_i} |y_i(p) - \hat{y}_i(p)| \,. \tag{3.19}$$

Masked mean squared error (MSE). The masked MSE emphasizes larger discrepancies by squaring the error before averaging:

$$\mathrm{MSE}_i = \frac{1}{|m_i|} \sum_{p \in m_i} (y_i(p) - \hat{y}_i(p))^2 \,. \tag{3.20}$$

Masked peak signal-to-noise ratio (PSNR). PSNR summarizes the reconstruction fidelity on a logarithmic scale derived from the masked MSE:

$$\mathrm{PSNR}_i = 10 \log_{10}\left(\frac{L^2}{\mathrm{MSE}_i}\right), \tag{3.21}$$

where L is the HU intensity range. In our setting, we use $L = 1200 - (-1024) = 2224$ HU. Higher PSNR indicates better agreement (i.e., lower MSE).

Masked structural similarity (SSIM). For two image patches (7×7 pixels) centered at pixel p with means $\mu_y, \mu_{\hat{y}}$, variances $\sigma_y^2, \sigma_{\hat{y}}^2$, and covariance $\sigma_{y\hat{y}}$, the local SSIM is defined as:

$$\mathrm{SSIM}(p) = \frac{(2\mu_y\mu_{\hat{y}} + c_1)(2\sigma_{y\hat{y}} + c_2)}{(\mu_y^2 + \mu_{\hat{y}}^2 + c_1)(\sigma_y^2 + \sigma_{\hat{y}}^2 + c_2)}, \tag{3.22}$$

where c_1 and c_2 are stability constants. The masked SSIM for slice i is then the mean of the local SSIM values over the body mask:

$$\mathrm{SSIM}_i = \frac{1}{|m_i|} \sum_{p \in m_i} \mathrm{SSIM}(p). \tag{3.23}$$

3.6 Dosimetric Evaluation

Image similarity metrics measure pixel-wise or perceptual agreement but do not directly capture whether intensity errors translate into clinically relevant dose calculation errors. In radiotherapy, even small systematic errors in electron density estimation-particularly at bone or air-tissue interfaces-can propagate into dose discrepancies that affect treatment outcomes. Dosimetric evaluation addresses this limitation by simulating actual treatment planning workflows and comparing the resulting dose distributions between reference CT and sCT (Chen et al., 2014).

In radiotherapy treatment planning, the irradiation geometry is defined relative to two complementary structure types. The *planning target volume* (PTV) is the region intended to receive the prescribed radiation dose, encompassing the tumor and margins for geometric uncertainty. *Organs at risk* (OARs) are healthy structures that must be spared from excessive radiation to minimize side effects. Treatment plan optimization seeks to maximize dose conformity to the PTV while simultaneously limiting dose to OARs – a trade-off that is formalized through dose-volume constraints during fluence optimization (Drzymala et al., 1991).

The dose-volume histogram (DVH) is the standard tool for summarizing three-dimensional dose distributions in treatment plan evaluation. A cumulative DVH curve expresses, for each dose level d, the fraction of a structure's volume receiving at least that dose. From this curve, several scalar metrics are derived to characterize plan quality:

Standard scalar metrics derived from the DVH include the maximum dose $D_{\max}$, the mean dose D_{mean}, and the family of D_x metrics, defined as the minimum dose received by the hottest $x\%$ of the structure volume. Clinically relevant instances include D_{95} and D_{98}, which characterize target coverage, and D_2, which serves as a near-maximum hot-spot indicator for OARs (International Commission on Radiation Units and Measurements, 2010).

Chapter 4

Data

This work uses paired MRI-CT image volumes provided by the SynthRAD 2023 and 2025 Grand Challenges, which were introduced in the context of the broader challenge design in Section 2.5. The present chapter describes the composition, acquisition characteristics, and preprocessing of the combined dataset in detail, with particular emphasis on the sources of heterogeneity that motivate the normalization strategies discussed in Section 3.1.

4.1 Dataset Origin and Composition

Both SynthRAD challenges were organized to benchmark sCT generation algorithms for radiotherapy planning. Each challenge released publicly available, anonymized datasets of rigidly registered MRI-CT pairs (Task 1) alongside CBCT-CT pairs (Task 2). Because this work focuses exclusively on MRI-to-CT synthesis, only Task 1 data from both challenges is used.

The SynthRAD2023 dataset was collected at three Dutch university medical centers – UMC Utrecht, UMC Groningen, and Radboud UMC Nijmegen – and covers two anatomical regions: brain and pelvis. It comprises 540 MRI-CT pairs in total, split into 180 training, 30 validation, and 60 test cases per region (Thummerer et al., 2023). However, only the training set (360 pairs) was fully released with ground-truth CTs; the validation and test sets were withheld for the challenge evaluation and are not publicly available with paired CT data. SynthRAD2025 substantially expanded the scope, collecting data from five European university medical centers: the three Dutch centers from the 2023 edition plus LMU Hospital Munich and the University Hospital of Cologne in Germany. It covers three different anatomical regions – head-and-neck, thorax, and abdomen – and comprises 890 MRI-CT pairs for Task 1 (Thummerer et al., 2025). As with SynthRAD2023, only the training partition (578 pairs) includes ground-truth CTs; validation and test ground truths are reserved for challenge evaluation.

As a result, all experiments in this work are conducted exclusively on the publicly available training partitions of both challenges. The resulting combined dataset comprises 938 patients across five anatomical regions, as detailed in Table 4.1, with representative axial slices for each anatomical region and center shown in Figure 4.1.

For anonymity, the dataset papers refer to the data-providing institutions using letter codes (A through D, center E only provided data to Task 2 and is therefore not relevant for this work) without disclosing the mapping to specific centers. This convention is adopted throughout this work. Not all centers contributed data to every anatomical region: for instance, head-and-neck MRI was provided by centers A, C, and D; thorax MRI by centers A and B; and abdomen MRI by centers A, B, and C.

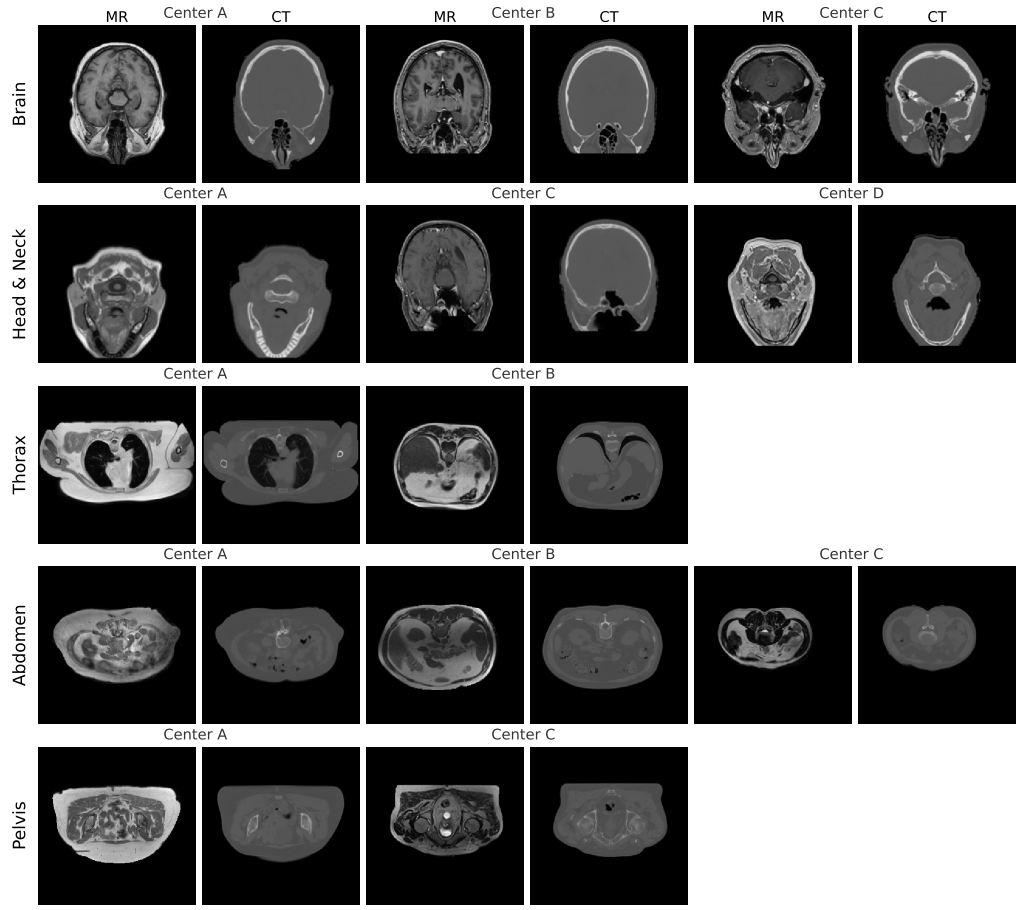


Figure 4.1: REPRESENTATIVE AXIAL SLICES FROM THE COMBINED DATASET. One randomly selected image slice per center and anatomical region is shown. Each pair shows the MR image (left) and its registered CT ground truth (right).

Table 4.1: COMPLETE DATASET COMPOSITION ACROSS SYNTHRAD CHALLENGES. Training set patient counts per anatomical region and contributing center. Center labels are independently anonymized per challenge and are not comparable across years.

SynthRAD2023

Region	Center A	Center B	Center C	Total
Brain	60	60	60	180
Pelvis	120	–	60	180
Total	**180**	**60**	**120**	**360**

SynthRAD2025

Region	Center A	Center B	Center C	Center D	Total
Head & Neck	91	–	65	65	221
Thorax	91	91	–	–	182
Abdomen	65	91	19	–	175
Total	**247**	**182**	**84**	**65**	**578**

4.2 Image Acquisition

The MR and CT images in the combined dataset were acquired under markedly different conditions. While CT protocols are largely standardized across centers by virtue of the HU scale, MRI acquisition varied substantially in scanner hardware, field strength, pulse sequence, and contrast agent usage – differences that are central to the normalization challenge addressed in this work.

4.2.1 MRI Acquisition

Across both challenges, MR images were acquired using scanners from three manufacturers – Philips, Siemens, and ViewRay – at field strengths ranging from 0.35 T to 3 T. A variety of pulse sequences were used, predominantly T1-weighted gradient echo and spin echo variants, but also T2-weighted and balanced steady-state free-precession (bSSFP) sequences. Notably, Center B acquired all its MRI data on a ViewRay MRidian MR-Linac at 0.35 T using a bSSFP sequence, which produces fundamentally different tissue contrast compared to the T1-weighted sequences used at the remaining sites (Thummerer et al., 2025).

The use of gadolinium-based contrast agents further compounds the variability. In SynthRAD 2023, brain MRIs from two of three centers were acquired with contrast enhancement, whereas pelvis MRIs were acquired without contrast at all centers (Thummerer et al., 2023). In SynthRAD2025, head-and-neck MRIs from all contributing centers were contrast-enhanced, while thorax and most abdomen acquisitions were not, with the exception of a subset of abdomen cases from Center A where roughly 30% received contrast (Thummerer et al., 2025). This means that the presence or absence of contrast varies not only between anatomical regions but also within a single region across centers and even within a single center.

Voxel spacings varied considerably across centers and regions. In-plane resolutions ranged from approximately 0.5 mm to 1.6 mm, while slice thicknesses ranged from 0.9 mm (isotropic acquisitions for head-and-neck at Center D) to 7.0 mm (some abdomen acquisitions at Center A). Both 2D and 3D acquisition modes were used, with some centers acquiring images under free breathing and others using breath-hold protocols, particularly relevant for thorax and abdomen regions where respiratory motion can cause artifacts (Thummerer et al., 2025).

4.2.2 CT Acquisition

In contrast to MRI, CT images were acquired with broadly consistent protocols across centers. All planning CTs used tube voltages of 120 kVp (with rare exceptions at 90-140 kVp for specific abdomen cases), and images were reconstructed on 512×512 grids. Pixel spacings ranged from approximately 0.6 mm to 1.5 mm, and slice thicknesses from 1 mm to 5 mm depending on the anatomical region and center. CT scanners from Philips, Siemens, Toshiba, and GE were used (Thummerer et al., 2023, 2025).

4.3 Preprocessing by Challenge Organizers

Before release, all datasets underwent a standardized preprocessing pipeline applied by the challenge organizers. This pipeline was designed to harmonize image geometry, ensure patient anonymity, and provide evaluation-ready data, while deliberately preserving the intensity variability inherent in multi-center MRI acquisitions (Thummerer et al., 2023, 2025).

The preprocessing consisted of several steps. First, all images were converted from DICOM to compressed volumetric file formats: NIfTI (`.nii.gz`) for SynthRAD2023 and MetaImage (`.mha`) for SynthRAD2025. MRI and CT volumes were then rigidly registered using the Elastix framework to align the paired images into a common coordinate system (Thummerer et al., 2023, 2025).

All images were resampled to uniform voxel grids: $1\times1\times1\,\text{mm}^3$ for brain, $1\times1\times2.5\,\text{mm}^3$ for pelvis (SynthRAD2023), and $1\times1\times3\,\text{mm}^3$ for all regions in SynthRAD2025. To ensure patient anonymity, brain and head-and-neck images were defaced by overwriting facial structures with background intensity values (−1024 HU for CT, 0 for MRI). In SynthRAD2023, defacing was performed manually based on eye contours; in SynthRAD2025, an automated algorithm based on TotalSegmentator skull and brain segmentations was used (Thummerer et al., 2023, 2025).

A binary patient outline mask was generated for each case using histogram-based thresholding and morphological operations. These masks define the evaluation region and were dilated to include a margin of surrounding air. The challenge organizers noted that variations in thresholding parameters across centers and patients can result in variable dilation margins, occasional inclusion of couch structures, or exclusion of peripheral anatomy such as lung regions (Thummerer et al., 2025).

Finally, all volumes were cropped to the bounding box of the patient outline mask with a margin of 20 voxels (SynthRAD2023) or 10 voxels (SynthRAD2025) to reduce file size. Crucially, no intensity normalization or harmonization was applied to the MRI data during preprocessing, preserving the full extent of inter-scanner and inter-protocol variability in the released data (Thummerer et al., 2023).

4.4 Data Heterogeneity

Given the acquisition diversity described above, identical anatomical structures can appear with substantially different intensity distributions depending on the acquisition setting. Figure 4.2 illustrates this for abdomen patients: while CT intensity distributions are largely consistent across centers due to the standardized Hounsfield scale, MRI intensity distributions differ markedly between centers, reflecting the scanner- and protocol-dependent nature of MRI signal formation. This distributional shift is not limited to a simple global scaling – it affects different tissue types to varying degrees, as the relative contrast between tissues depends on the specific combination of field strength, sequence, and timing parameters.

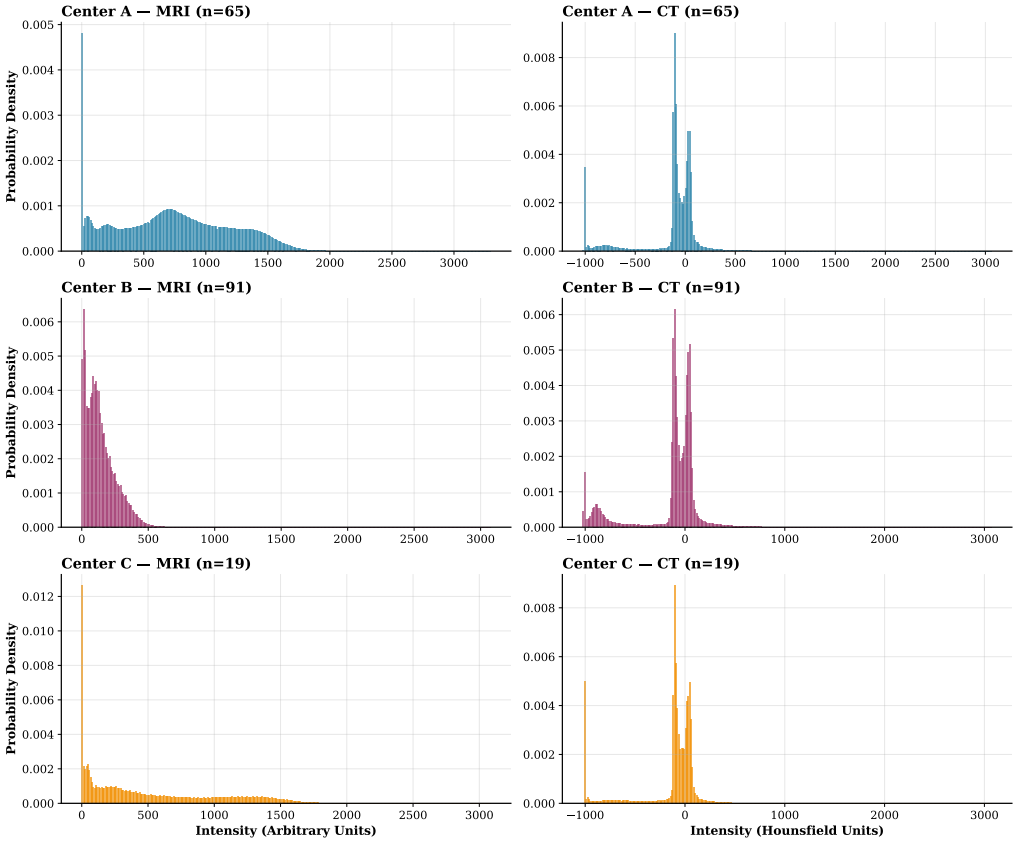


Figure 4.2: INTER-CENTER COMPARISON OF VOXEL INTENSITY DISTRIBUTIONS FOR ABDOMEN PATIENTS. CT intensities (Hounsfield Units) are standardized and comparable across centers, whereas MRI intensities show substantial distributional shifts, reflecting scanner- and protocol-dependent acquisition differences.

4.5 Train/Validation/Test Split

All experiments are conducted exclusively on the fully accessible data provided by the SynthRAD2023 and SynthRAD2025 challenges. Table 4.1 summarizes the complete distribution across years, anatomical regions, and centers. A stratified split was performed to preserve the relative distribution across challenge year, anatomical region, and acquisition center, following an approximately 70/15/15 ratio. Table 4.2 confirms that this balance holds at the combined region and center level.

Table 4.2: FINAL TRAIN/VALIDATION/TEST DISTRIBUTION AT BODY AND CENTER LEVEL.

Body	Center	Train	Val	Test	Total
Abdomen	A	45	10	10	65
Abdomen	B	64	13	14	91
Abdomen	C	13	3	3	19
Brain	A	42	9	9	60
Brain	B	42	9	9	60
Brain	C	42	9	9	60
Head-Neck	A	64	14	13	91
Head-Neck	C	45	10	10	65
Head-Neck	D	45	10	10	65
Pelvis	A	84	18	18	120
Pelvis	C	42	9	9	60
Thorax	A	64	13	14	91
Thorax	B	64	14	13	91
Total		**656**	**141**	**141**	**938**

Chapter 5

Pipeline and Approach

This chapter describes the complete methodological pipeline developed for sCT generation from MR images. Building upon the theoretical foundations introduced in the background (chapter 3), we integrate dataset harmonization, geometry standardization, domain-aware modeling strategies, transformer-based generator architectures, and clinically motivated evaluation procedures into a unified framework. Figure 5.1 summarizes the individual building blocks and their interactions, from raw multi-center data to quantitative and dosimetric assessment.

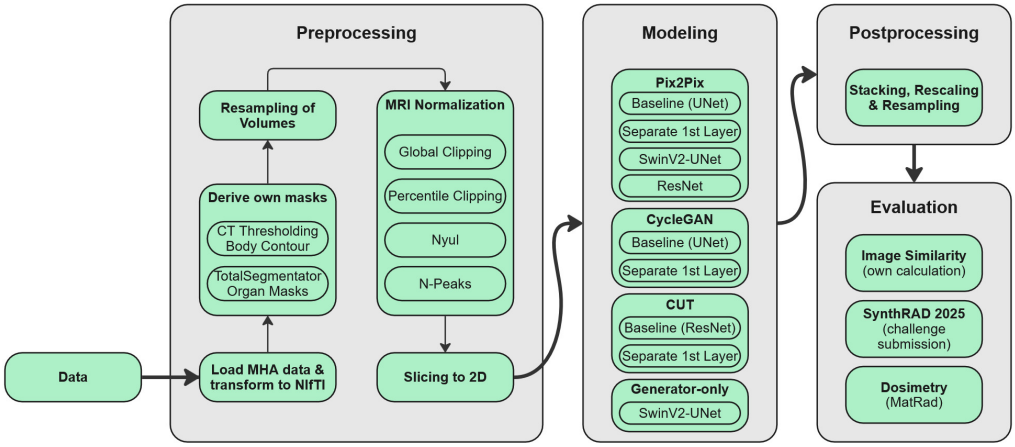


Figure 5.1: Overview of the MR-to-CT Synthesis Pipeline. The building blocks of the pipeline, covering preprocessing (resampling, MRI normalization, masking, and slicing), modeling (Pix2Pix, CycleGAN, CUT, and generator-only variants), postprocessing (stacking, rescaling, and resampling), and evaluation (image similarity, SynthRAD2025 submission, and dosimetry).

5.1 Preprocessing

This section describes the preprocessing steps applied to the MR and CT data prior to model training. Depending on the specific experiment, not all components of the preprocessing pipeline are utilized, as certain steps are selectively enabled to isolate methodological effects.

5.1.1 Derivation of Masks

Three types of masks are derived for distinct purposes throughout the pipeline: CT-threshold-based body contour masks, MR-based organ delineations, and CT-based organ delineations.

CT-Threshold-Based Body Contour Masks. Because the provided SynthRAD body masks are overly broad, we derive tighter body contours directly from the registered CT to improve training precision without relying on external scanning devices. For each CT volume, voxels are first thresholded at $-400\,\mathrm{HU}$ to separate body from air. The resulting binary mask is cleaned via two rounds of erosion, after which the largest external contour is extracted slice-wise in the axial plane to retain only the main body outline and remove small disconnected regions such as arms. Coverage is then restored with five rounds of dilation, followed by binary closing to seal gaps – notably around the nose and air cavities in head-and-neck volumes – and a final 3D hole-filling step. The resulting mask is saved as NIfTI using the original affine.

MR-Based Organ Delineation. TotalSegmentator is used to extract masks as MR-guided tissue priors for intensity standardization. The following masks are extracted: *liver, torso_fat, subcutaneous_fat, skeletal_muscle*. These masks are then used in the N-peaks workflow (Section 3.1.2) to estimate tissue-specific histogram peaks and normalize MR intensities consistently across patients.

CT-Based Organ Delineation. For dosimetric evaluation, TotalSegmentator masks are generated directly on the real CT volumes. These CT-based masks serve as structure definitions for the DICOM RT Structure Set (RTSTRUCT) file required by the dose-volume histogram (DVH) analysis. The CT is used as the source because the *skin* contour is needed and TotalSegmentator does not yet offer a skin segmentation task for MRI.

5.1.2 Volume Resampling

The goal of resampling is bringing all MR and CT volumes into consistent geometries. First, the anatomy is localized by cropping to the body-mask bounding box along the two axial (in-plane) axes, while the third axis (number of slices) is left unchanged. This ensures that the next step does not crop tissue only because the body mask was not at the center of the volume.

For thorax, abdomen, and pelvis scans, which have larger original fields of view, we first standardize to 512×512 in-plane by padding or cropping (depending on the dimensions of the respective volume), then downsampling by a factor of two to obtain the final 256×256 resolution, effectively doubling the pixel spacing. No interpolation is performed during this step, as the downsampling factor is an integer divisor of the grid. For head-and-neck and brain scans, which have smaller original fields of view, we directly crop to 256×256 without any resampling.

Because no resampling is applied to the head-and-neck and brain volumes, these retain their original (finer) pixel spacing and therefore represent anatomical structures at roughly twice the spatial resolution per axis compared to the body scans. This trade-off is accepted in favor of maximizing information in input images while minimizing the amount of cropped tissue. Appendix A.1 illustrates the original mask size distributions and the effect of this standardization.

After geometric harmonization, voxels outside the body mask are set to modality-specific background values ($-1000\,\mathrm{HU}$ for CT, 0 for MR), such that only the relevant anatomy remains. The final outputs are aligned MR/CT volumes and matching masks with consistent axial dimensions, ready for downstream normalization, slicing, and training.

5.1.3 Data Standardization

Several types of normalization are utilized across the experiments. CT images are always normalized with the same approach, MR images are normalized with one of the options described in the respective sections.

CT: Baseline Normalization

CT images (denoted as Y' before and Y after standardization) are normalized by clipping to a fixed HU window and applying min-max scaling:

$$Y = \begin{cases} 0 & \text{if } Y' < -1024 \\ \dfrac{Y' + 1024}{2224} & \text{if } -1024 \leq Y' \leq 1200 \\ 1 & \text{if } Y' > 1200 \end{cases} \tag{5.1}$$

This maps air (-1024 HU) to 0 and dense bone (1200 HU) to 1, while suppressing implant or noise outliers beyond this range. CT normalization is kept identical across all experiments.

For MRI (denoted as X' before and X after standardization), four normalization strategies of increasing sophistication are compared.

MRI: Baseline Normalization

The simplest strategy applies a fixed global clipping threshold to suppress extreme high-intensity outliers, followed by min-max scaling to $[0, 1]$:

$$X = \begin{cases} \dfrac{X'}{2000} & \text{if } 0 \leq X' \leq 2000 \\ 1 & \text{if } X' > 2000 \end{cases} \tag{5.2}$$

mapping the background to 0 and the clipped maximum to 1. While this removes gross outliers, it does not account for scanner- or patient-specific intensity distributions.

MRI: Per-Image P99 Normalization

As an adaptive alternative, we evaluate normalization based on the 99^{th} percentile of per-volume foreground intensities. The 99^{th} percentile (p_{99}) is computed over non-zero voxels to exclude the background, and all intensities are divided by this value and clipped to $[0, 1]$:

$$X = \begin{cases} \dfrac{X'}{p_{99}} & \text{if } 0 \leq X' < p_{99} \\ 1 & \text{otherwise} \end{cases} \tag{5.3}$$

Using the 99^{th} percentile rather than the global maximum reduces the influence of bright outliers such as hyperintense artifacts or contrast-enhanced vessels, while the per-image computation compensates for scan-to-scan variability without requiring tissue masks or external reference data.

MRI: Nyúl and N-Peaks Normalization

Beyond these global approaches, two histogram-based methods are evaluated that explicitly align tissue intensity distributions across subjects. Nyúl histogram matching learns a standard scale from the training set and maps each image's intensity landmarks to it via piecewise linear interpolation, reducing inter-scanner variability at the cohort level. N-Peaks extends this principle by anchoring the transformation to tissue-specific histogram peaks derived from anatomically defined regions, making it less sensitive to inter-patient differences in body composition. Detailed descriptions of both methods, including their mathematical formulations, are provided in Section 3.1.

For the abdominal region, N-Peaks is configured with liver and a combined fat mask (merging TotalSegmentator's `torso_fat` and `subcutaneous_fat` labels) as tissue anchors, together with the image background fixed at zero intensity. Liver provides a landmark in the mid-intensity range of T1-weighted acquisitions, while the combined fat mask anchors the high-intensity range. When N4 bias field correction is applied prior to N-Peaks normalization, it uses the default settings of 5 fitting levels with 75 iterations per level and a convergence threshold of 0.01. Reducing the iterations per level to 15 produced visually indistinguishable results, so the default setting is retained.

5.1.4 Slicing to 2D

Since all investigated architectures are formulated as 2D models, the preprocessed 3D MR and CT volumes are decomposed into axial 2D slices. For each patient, spatially aligned MR and CT volumes (together with the corresponding body mask) are loaded as 3D tensors of shape (H, W, D), where H, W, and D denote height, width, and depth, respectively. We then iterate along the axial dimension z and extract individual slices, resulting in 2D samples of size $(H, W, 1)$. This procedure transforms each volumetric sample into a set of independent 2D training pairs, enabling the use of 2D generative models while maintaining intra-slice spatial consistency between MR and CT domains.

5.2 Modeling

The goal of this stage is learning a mapping $G : \mathcal{X} \to \mathcal{Y}$ where, in this project, $\mathcal{X}$ denotes the MR image domain and $\mathcal{Y}$ the CT image domain. Since both domains share the same underlying anatomical structures, but differ in intensity distribution and imaging physics, the generator must simultaneously preserve spatial consistency and model complex appearance transformations. Architecturally, this requires (i) a sufficiently large receptive field to capture global context, and (ii) mechanisms to retain high-resolution spatial information. Encoder-decoder designs with skip or residual connections have proven particularly effective in this setting, as they allow low-level structural information to bypass the bottleneck while deeper layers model high-level transformations.

While the core image-to-image translation architectures – Pix2Pix (Section 3.3.1), CycleGAN (Section 3.3.2), and CUT (Section 3.3.3) – as well as the backbone networks UNet-256 (Section 3.4.1) and ResNet-9 (Section 3.4.2) were described in their respective sections, we extend these baseline configurations in two directions: first, treatment-center-specific input modeling via a *separate first layer*, and second, replacement of the convolutional backbones by a transformer-based *SwinV2-UNet*.

5.2.1 Separate First Layer for Domain-Specific Adaptation

As described in Section 4.4, MR images originate from multiple treatment centers and scanner configurations, resulting in residual domain shifts despite normalization. Instead of relying exclusively on manual preprocessing to harmonize intensities, we also experiment with a model-based alternative referred to as *separate first layer*.

Concretely, the first convolutional layer of the respective generator (UNet or ResNet) is replaced by parallel convolutional layers of identical architecture and dimensionality, one per treatment center $d \in \{A, B, C\}$. Given an input slice, only the convolution corresponding to its center is activated, while the remaining two are ignored. The modified first layer computes

$$h = \mathcal{C}_d(x), \tag{5.4}$$

where $\mathcal{C}_d$ represents the convolution associated with center d for input image x. All subsequent layers are shared across domains.

This design can be interpreted as a shallow, domain-specific feature extractor followed by a fully shared generator. The approach preserves parameter efficiency compared to training independent models per center, while allowing low-level filters (e.g. intensity and texture-sensitive kernels) to specialize to scanner-specific characteristics. Since intensity distributions and noise patterns primarily affect early convolutional responses, restricting domain adaptation to the first layer provides a targeted mechanism to mitigate inter-center variability without altering the overall architecture. Also, the number of parameters of the first layer is lower than of the subsequent layers, which increases the likelihood of successful adaptation despite limited data.

5.2.2 SwinV2-UNet Generator

To investigate whether improved global context modeling benefits MR-to-CT translation, we replace convolutional generators with a transformer-based alternative Swin Transformer (see Section 2.1.1). Specifically, we implement a *SwinV2-UNet*, structurally analogous to Swin-UNet (Cao et al., 2023), but employing *Swin Transformer V2* blocks (Liu et al., 2022) instead of the original Swin backbone (Liu et al., 2021).

The hierarchical design produces multi-scale feature maps compatible with the UNet paradigm and with the input resolution used in this project (256^2 instead of 224^2 as for SwinV1).

The architecture retains the U-shaped encoder-decoder topology with skip connections, ensuring comparability to UNet-based generators. The encoder consists of hierarchical SwinV2 stages with patch merging operations that progressively reduce spatial resolution while increasing embedding dimensionality. The decoder mirrors the encoder structure using patch expanding layers to progressively restore spatial resolution while reducing embedding dimensionality. Skip connections fuse encoder and decoder representations at matching scales. Furthermore, instead of the final linear projection layer used in Swin-UNet, we employ a 1×1 convolution followed by a non-linear output activation constraining the range of the final intensity prediction. By doing so, we port the model from segmentation tasks to image-to-image translation.

We employ the SwinV2-Tiny configuration as backbone, which provides a favorable trade-off between model capacity and data efficiency. The ImageNet-pretrained model is used to initialize the encoder (including the *bottleneck*), except for the first patch embedding layer: originally, this layer partitioned an RGB image into non-overlapping 4×4 patches, forming $4 \times 4 \times 3 = 48$-dimensional vectors that are linearly projected to an embedding dimension of 96. For our single-channel MR-to-CT setting, it is replaced by mapping $4 \times 4 \times 1 = 16$-dimensional patches to 96 channels, resulting in $1{,}632$ randomly initialized, learnable parameters and resolving the channel mismatch with pretrained weights. The decoder is initialized randomly, and we evaluate two

settings: (i) fully trainable encoder and decoder, and (ii) a frozen encoder (except for the patch embedding layer). The full architecture is depicted in Figure 5.2.

In contrast to convolutional generators, the SwinV2-UNet relies on self-attention within local windows and shifted windows, enabling modeling of long-range spatial dependencies beyond the effective receptive field of standard convolutions. This property is particularly relevant for MR-to-CT translation, where bone structures and air cavities may require global anatomical context for accurate intensity prediction.

5.2.3 Generator-only Setting

To disentangle the effect of adversarial training from the representational capacity of the generator architecture itself, we introduce a *generator-only* setting. This configuration is exclusively applied to the SwinV2-UNet described above and removes the discriminator entirely, thereby excluding the cGAN objective.

In the standard Pix2Pix formulation (3.9), the generator G is optimized with respect to a combined objective consisting of the adversarial loss $\mathcal{L}_{\text{cGAN}}(G, D_Y)$ (3.10) and the pixel-wise reconstruction loss $\mathcal{L}_{L_1}(G)$ (3.11). In contrast, the generator-only setting optimizes G solely based on the L_1 reconstruction term:

$$G^* = \arg\min_G \mathcal{L}_{L_1}(G). \tag{5.5}$$

Accordingly, no discriminator D_Y is instantiated, and no adversarial gradients are propagated during training. The model therefore performs direct supervised regression from MR to CT, learning a deterministic mapping that minimizes the expected absolute deviation between predicted and reference CT intensities.

This configuration serves two purposes. First, it provides a controlled baseline to assess whether improvements observed in adversarial settings are attributable to the GAN objective or to the underlying SwinV2-UNet architecture. Second, it enables evaluation of whether the L_1 loss alone is sufficient for clinically relevant MR-to-CT translation, particularly with respect to low-frequency anatomical consistency and quantitative intensity accuracy.

By using the identical reconstruction loss as in the Pix2Pix formulation, differences in performance could be attributed to the presence or absence of adversarial training rather than to changes in optimization targets.

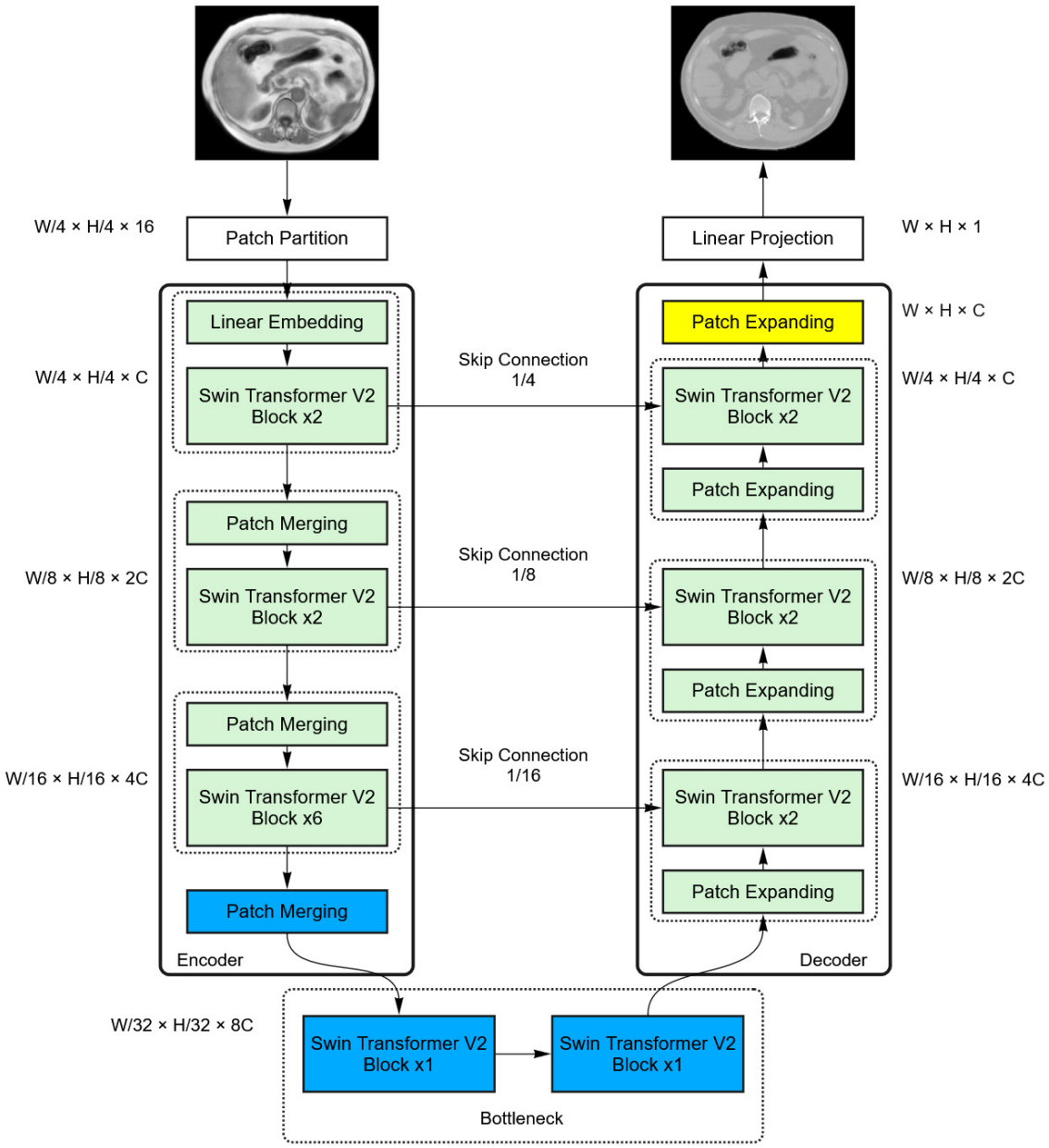


Figure 5.2: SWINV2-UNET ARCHITECTURE FOR MR-TO-CT SYNTHESIS. Our SwinV2-UNet architecture based on SwinV2-Tiny, inspired by Cao et al. (2023). In our SwinV2-UNet, the patch partitioning has 16 instead of 48 channels (due to grayscale MR inputs), the original Swin Transformer blocks are replaced by Swin Transformer V2 blocks, and the final linear projection layer is substituted with a 1×1 convolution followed by a `tanh` activation function.

5.3 Postprocessing

The trained translation models produce 2D sCT slices in the network output space (normalized intensity range). Postprocessing therefore serves two purposes: (i) mapping model outputs into physically meaningful units (HU) for quantitative evaluation and visualization, and (ii) reconstructing patient-level 3D sCT volumes, geometrically consistent with the original reference data.

Intensity mapping to Hounsfield units. Model outputs are first converted from the network's normalized range $(-1, 1)$ to HU using a fixed linear rescaling between bounds $(-1024, 1200)$:

$$\hat{\mathrm{X}}' \;=\; 2224 \cdot \frac{(\hat{X} + 1)}{2} \;-\; 1024 \tag{5.6}$$

Masking and slice-wise evaluation. A binary body mask (per-slice) is used to compute masked image quality metrics (MAE, MSE, PSNR, SSIM, see Section 3.5). Postprocessed slices are exported as NIfTI for inspection and downstream processing, alongside per-slice metrics written to tabular form (including the number of mask voxels).

3D volume reconstruction at original dimensions. For patient-level analysis, 2D sCT NIfTI slices are grouped by patient identifier and stacked in slice-index order to form a resampled-resolution 3D volume. The reconstruction restores a consistent anatomical orientation. A reference affine transformation (from input MR) is used to attach spatial metadata to the reconstructed volume. To obtain an sCT volume matching the *original* patient-specific dimensions, a reverse-resampling step is applied (c.f. resampling in Section 5.1.2): for body regions that have been downsampled to the canonical in-plane resolution, the reconstructed volume is upsampled in-plane and then centered into the target field-of-view via symmetric padding/cropping, while preserving the slice direction. Voxels outside the body mask (in canonical space) are assigned an air-equivalent value of -1024 HU prior to reversing the resampling, ensuring a well-defined background in the restored geometry.

5.4 Evaluation

Several types of evaluations are implemented for different purposes.

Volume-level aggregation. To report a single score per 3D volume, we aggregate slice-wise metrics using mask-size weighting so that slices with larger body regions contribute proportionally more:

$$\mathrm{Metric}_{\mathrm{vol}} = \frac{\sum_i |m_i| \cdot \mathrm{Metric}_i}{\sum_i |m_i|}. \tag{5.7}$$

where m_i denotes the binary foreground mask of slice i and $|m_i|$ its number of non-zero voxels. This aggregation is applied consistently to the image similarity metrics MAE, MSE, PSNR, and SSIM (see Section 3.5), producing the reported values.

5.4.1 Evaluation within SynthRAD2025

The SynthRAD2025 challenge allows submission of algorithms in a dedicated *post-challenge* phase. These submissions are evaluated using the same protocol as during the official challenge but are

reported on a separate leaderboard. The evaluation is carried out independently by the challenge organizers based on a private test set and the results can directly be compared to other submissions.

The evaluation comprises three categories of metrics (the detailed definitions and descriptions of the exact computation and ranking scheme can be found on the challenge webpage[1]):

- **Image similarity:** mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and multi-scale structural similarity index (MS-SSIM).
- **Geometric consistency:** multiclass Dice coefficient (mDice) and 95th percentile Hausdorff distance (HD95).
- **Dose evaluation:** mean absolute dose difference, dose-volume histogram (DVH) metrics, and gamma pass rates (GPR), each computed separately for photon and proton treatment plans.

The metrics are evaluated within the body contour and compare the algorithm output to the ground-truth CT.

5.4.2 Dosimetric evaluation using matRad

To assess the clinical relevance of the generated sCT images beyond voxel-wise similarity metrics, we implement a dosimetric evaluation pipeline using the open-source treatment planning toolkit matRad (Wieser et al., 2017).

The dosimetric evaluation pipeline consists of four main stages:

Stage 1: Format Conversion. Since matRad's dosimetric pipeline requires DICOM (Digital Imaging and Communications in Medicine) files as input, the NIfTI volumes (CT and sCT) are converted to DICOM format prior to treatment planning. Voxel spacing, orientation, and origin metadata are preserved during conversion to ensure geometric consistency across all subsequent stages.

Stage 2: Structure Definition. Anatomical structures required for treatment planning are automatically segmented from the reference CT using TotalSegmentator. For the abdominal test cases, the following structures are extracted: liver, left kidney, right kidney, spinal cord, and skin (body contour). These binary masks are also converted to DICOM format. The left kidney is designated as the planning target volume (PTV) – the region intended to receive the prescribed radiation dose – while the remaining structures serve as organs at risk (OARs), which should be spared from excessive radiation to minimize treatment side effects.

Stage 3: Treatment Plan Optimization. Treatment planning is performed using matRad's photon dose calculation engine with a generic linear accelerator model. The planning configuration employs 15 gantry angles with a bixel (discrete beam pixels subdividing the radiation field) width of 4 mm and a dose grid resolution of $4 \times 4 \times 4\,\mathrm{mm}^3$. The optimization objective for the target is defined as a minimum DVH constraint requiring at least 95% of the target volume to receive 54 Gy, while OARs are assigned mean dose objectives with a constraint of 15 Gy. Fluence optimization is performed on the reference CT, yielding optimized beam weights $\mathbf{w}^*$.

Stage 4: Dose Recalculation and DVH Analysis. Using the beam weights $\mathbf{w}^*$ optimized on the reference CT, forward dose calculation is performed on both the reference CT and the corresponding sCT. While the current clinical standard optimizes the treatment plan directly on the sCT and re-evaluates on the reference CT to assess plan transferability in an MRI-only workflow (Largent et al., 2019), this work intentionally adopts the reverse approach. By fixing $\mathbf{w}^*$ and recalculating dose on the sCT, any discrepancy in the resulting dose distribution can be attributed

[1] https://synthrad2025.grand-challenge.org/metrics-ranking/

solely to HU differences between the two images, providing a direct and unconfounded quantification of how intensity errors in the sCT propagate into dosimetric errors. The resulting dose distributions are analyzed using matRad's DVH computation routines, and delta metrics are computed as $\Delta = \text{sCT} - \text{CT}$ to quantify systematic deviations.

For each structure, DVH-derived metrics including mean dose, maximum dose ($D_{\max}$), D_{95}, D_{98}, and D_2 are extracted for both CT and sCT, and delta values are computed to quantify systematic deviations. Due to computational and licensing constraints, the evaluation is performed only on three abdominal test images (AB_1ABA068, AB_1ABB070, AB_1ABC127).

Chapter 6

Experiments

This chapter describes the experimental design used to evaluate different image-to-image translation frameworks for sCT generation from MR images. The goal is to systematically assess the influence of (i) data composition (joint *fullbody* training versus region-specific training), (ii) translation framework (Pix2Pix, CycleGAN, CUT), (iii) input MR image normalization techniques, and (iv) architectural choices (e.g. generator networks) on image similarity and downstream clinical relevance.

Figure 6.1 provides an overview of the complete experimental pipeline (details of the respective building blocks are described in Chapter 5). Starting from the SynthRAD2023 and SynthRAD2025 datasets, volumes are preprocessed and sliced into 2D images. The resulting slices are used to train different image-to-image translation frameworks with various generator architectures. Postprocessing includes stacking slice-wise predictions into 3D volumes and resampling to the original geometry. Evaluation is performed using image similarity metrics, the official SynthRAD2025 post-challenge evaluation, and dosimetric assessment using matRad.

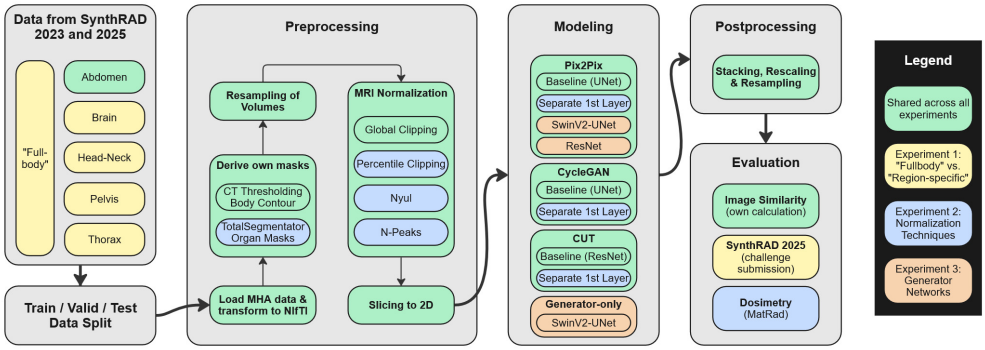


Figure 6.1: OVERVIEW OF THE EXPERIMENTAL PIPELINE. Overview of the building blocks of the experimental pipeline. The colors indicate for which experiment the block is relevant, as not all blocks were used for all experiments.

Across all experiments, the same building blocks for preprocessing, modeling, postprocessing, and evaluation are applied to ensure comparability. Differences between experiments are restricted to the investigated variable (e.g. normalization strategy or training regime) and are described in more detail in the respective following sections. All models have been trained and

evaluated on identical train/validation/test splits (see Section 4.5).

6.1 Experiment 1: Fullbody vs. Separate Models by Anatomical Region (RQ1)

The first experiment investigates whether a single model trained jointly on all anatomical regions ("fullbody") can achieve comparable performance to models trained separately for each region. This experiment addresses the following research question:

> *RQ1: Can a jointly trained fullbody model achieve comparable performance to region-specific models for MR-to-CT translation, under otherwise identical training conditions?*

The motivation for this comparison is twofold. First, a unified model simplifies deployment and maintenance in clinical workflows. Second, joint training may allow the model to exploit shared structural patterns across anatomical regions, especially as the images of different anatomical regions overlap at their intersections. However, anatomical heterogeneity may also hinder learning and reduce region-specific fidelity. This experiment aims to quantify this trade-off.

6.1.1 Setup

Data. The experiment is conducted on the SynthRAD datasets, including the anatomical regions abdomen, brain, head-and-neck, pelvis, and thorax. For the *fullbody* setting, data from all regions are pooled and used to train a single model. For the *region-specific* setting, five separate datasets are created, each containing only samples from one anatomical region. Preprocessing follows the pipeline described in Section 5.1, including resampling, body masking, baseline MRI normalization, and 2D axial slicing.

Model Frameworks. Three image-to-image translation frameworks are evaluated: *Pix2Pix* (cGAN), *CycleGAN* (unpaired GAN with cycle-consistency), and *CUT* (contrastive unpaired translation). For each framework, the baseline generator architecture is used (UNet for Pix2Pix and CycleGAN, ResNet-based generator for CUT). Details can be found in Section 5.2. A total of 18 models are trained and evaluated; for each of the three frameworks, one fullbody model and five region-specific models (one per anatomical region). All models are trained under identical optimization settings (learning rate 0.0002, batch size 1, number of epochs 50).[1] All training configurations are listed in Appendix Section A.2.

Postprocessing and Evaluation. Evaluation is performed per anatomical region to ensure fair comparison between the fullbody and region-specific models. All metrics are computed on the identical held-out test set defined in Section 4.5, ensuring patient-level separation between training and evaluation. Two types of evaluation are conducted, an image similarity-based analysis (see Section 3.5) as well as through submissions to the post-challenge phase of SynthRAD2025 (see Section 5.4.1).

[1] Attempts to achieve visibly higher performance by changing hyperparameters were not successful.

6.1.2 Results

The results are reported separately for each framework but aggregated over all anatomical regions. A full overview also broken down by body region can be found in appendix Section A.3.

The results of the image similarity evaluation using the voxel-wise error metrics MAE and MSE as well as the perceptual similarity measures PSNR and SSIM are computed for each anatomical region and subsequently aggregated over all five regions. Table 6.1 reports mean and standard deviation across all test cases, while the boxplots in Figure 6.2 illustrate the distribution of per-case performance and highlight potential outliers. No formal statistical significance testing is performed; therefore, differences should be interpreted descriptively, particularly in cases of overlapping standard deviations.

Table 6.1: Experiment 1 Results Aggregated Over All Anatomical Regions (abdomen, brain, head-neck, pelvis, thorax) based on own image similarity evaluation (mean ± std), test set size = 141 volumes. The best-performing model within each framework is highlighted in bold; the overall best model is additionally marked in italic.

Model	Type	MAE ↓	MSE ↓	PSNR ↑	SSIM ↑
CUT	Fullbody	176.53 ± 110.7	154'865 ± 175'794	**19.38 ± 2.52**	**0.669 ± 0.075**
CUT	Region	**174.98 ± 114.5**	**142'676 ± 178'861**	19.30 ± 2.72	0.666 ± 0.092
Pix2Pix	Fullbody	137.19 ± 59.92	77'537 ± 66'337	19.97 ± 1.86	0.687 ± 0.069
Pix2Pix	Region	***128.76 ± 51.25***	***68'364 ± 52'284***	***20.38 ± 1.97***	***0.697 ± 0.072***
CycleGAN	Fullbody	257.28 ± 89.68	184'653 ± 100'464	15.94 ± 2.80	**0.538 ± 0.085**
CycleGAN	Region	**230.37 ± 91.43**	**163'022 ± 106'882**	**16.70 ± 2.88**	0.536 ± 0.088

Across frameworks, Pix2Pix achieves the best overall image similarity. In this setting, region-specific training consistently outperforms the fullbody model in all four metrics, yielding lower MAE and MSE as well as higher PSNR and SSIM. For CUT, differences between fullbody and region-specific training are marginal, with nearly identical central tendencies and overlapping distributions. In contrast, CycleGAN exhibits a clearer gap, where region-specific models improve MAE, MSE, and PSNR compared to the joint model, although SSIM remains comparable. Overall, the own evaluation indicates that joint training does not lead to a systematic performance gain and that, depending on the framework, slight degradations can be observed when pooling all anatomical regions.

The region-wise analysis presented in Appendix Section A.3 confirms that these trends remain largely consistent when results are examined separately for each anatomical region.

The SynthRAD2025 post-challenge evaluation complements this analysis by providing standardized and independent image similarity, geometric consistency, and dosimetric metrics (Tables 6.2 and 6.3, additionally, Figure 6.3 for easier readability of overall trends). The independent testing procedure allows comparing the results to the challenge winners.

Consistent with the own evaluation, CUT and Pix2Pix clearly outperform CycleGAN across most metrics. Differences between fullbody and region-specific models remain small for CUT, with slightly better PSNR for the region-specific model but marginally improved MS-SSIM, DICE, and HD95 for the fullbody model. For Pix2Pix, the ranking between joint and separate training varies depending on the metric, without a consistent advantage for either setting. In the dosimetric analysis, CUT achieves the most favorable results overall, and the discrepancies between fullbody and region-specific training are minor compared to the performance gap to CycleGAN.

Importantly, all investigated configurations remain substantially worse than the official challenge winners in both image-based and dosimetric metrics, underlining the remaining performance gap to state-of-the-art methods. With respect to RQ1, the results indicate that jointly

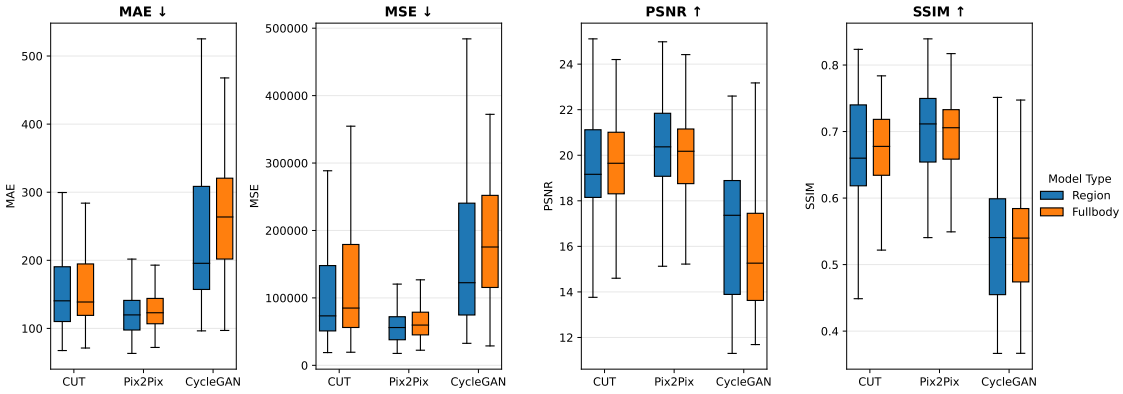


Figure 6.2: EXPERIMENT 1 RESULTS AGGREGATED OVER ALL ANATOMICAL REGIONS (abdomen, brain, head-neck, pelvis, thorax) based on own image similarity evaluation, test set size = 141 volumes.

Table 6.2: EXPERIMENT 1 RESULTS BASED ON SYNTHRAD2025 POST-CHALLENGE EVALUATION. Image similarity & geometric consistency metrics (mean ± std), post-challenge evaluation by SynthRAD2025 aggregated over *three* anatomical regions (abdomen, head-neck, thorax) on the private SynthRAD test set (223 volumes). Note that only three anatomical regions were included in the official post-challenge evaluation, which partially explains the deviation from the five-region aggregation used in the internal evaluation. The best-performing model within each framework is highlighted in bold; the overall best own model is additionally marked in italic.

Model	Type	MAE ↓	PSNR ↑	MS-SSIM ↑	DICE ↑	HD95 ↓
CUT	Fullbody	188.57 ± 49.33	21.90 ± 1.81	***0.72 ± 0.11***	***0.44 ± 0.15***	***17.89 ± 11.17***
CUT	Region	**185.90 ± 47.24**	***22.12 ± 1.81***	0.71 ± 0.11	0.42 ± 0.17	18.41 ± 10.81
Pix2Pix	Fullbody	***184.91 ± 51.09***	**21.96 ± 1.74**	**0.71 ± 0.10**	0.38 ± 0.12	22.15 ± 10.69
Pix2Pix	Region	187.40 ± 48.04	21.81 ± 1.63	0.70 ± 0.11	**0.40 ± 0.14**	**21.55 ± 11.15**
CycleGAN	Fullbody	259.77 ± 82.51	20.30 ± 2.70	0.52 ± 0.16	0.07 ± 0.09	**48.09 ± 63.94**
CycleGAN	Region	**221.63 ± 73.65**	**21.21 ± 2.63**	**0.58 ± 0.15**	**0.10 ± 0.11**	51.96 ± 56.11
Winners	Region	64.81 ± 21.25	30.00 ± 2.76	0.94 ± 0.05	0.78 ± 0.11	6.01 ± 3.24

Table 6.3: DOSIMETRIC EVALUATION METRICS BASED ON SYNTHRAD2025 POST-CHALLENGE EVALUATION. Dosimetric evaluation metrics (mean ± std), post-challenge evaluation by SynthRAD2025 aggregated over *three* anatomical regions (abdomen, head-neck, thorax) on the private SynthRAD test set (223 volumes). Note that only three anatomical regions were included in the official post-challenge evaluation, which explains the deviation from the five-region aggregation used in the internal evaluation. The best-performing model within each framework is highlighted in bold; the overall best own model is additionally marked in italic.

Model	Type	GPR 2mm/2% (photon) ↑	GPR 2mm/2% (proton) ↑	DVH error (photon) ↓	DVH error (proton) ↓	Dose MAE (photon) ↓	Dose MAE (proton) ↓
CUT	Fullbody	***96.76 ± 6.58***	***71.21 ± 13.38***	***0.21 ± 0.56***	***0.53 ± 1.25***	***0.010 ± 0.013***	***0.040 ± 0.068***
CUT	Region	96.18 ± 7.82	70.18 ± 11.70	0.22 ± 0.56	0.75 ± 3.86	0.012 ± 0.016	0.042 ± 0.068
Pix2Pix	Fullbody	**95.36 ± 8.61**	69.23 ± 12.84	0.22 ± 0.56	**4.82 ± 64.0**	**0.012 ± 0.016**	***0.040 ± 0.029***
Pix2Pix	Region	94.63 ± 8.34	**70.03 ± 13.50**	***0.21 ± 0.56***	11.96 ± 168.1	0.013 ± 0.015	0.044 ± 0.071
CycleGAN	Fullbody	78.42 ± 22.7	54.76 ± 17.85	0.41 ± 0.62	1.20 ± 6.41	0.045 ± 0.117	0.131 ± 0.267
CycleGAN	Region	**83.20 ± 22.5**	**59.69 ± 16.81**	**0.38 ± 0.62**	**1.02 ± 4.31**	**0.036 ± 0.098**	**0.125 ± 0.268**
Winners	Region	98.33 ± 5.41	84.04 ± 10.54	0.19 ± 0.56	0.35 ± 1.18	0.006 ± 0.009	0.024 ± 0.067

trained fullbody models retain performance comparable to region-specific models for CUT and Pix2Pix, whereas CycleGAN shows a clearer performance degradation under joint training, suggesting higher sensitivity to anatomical heterogeneity.

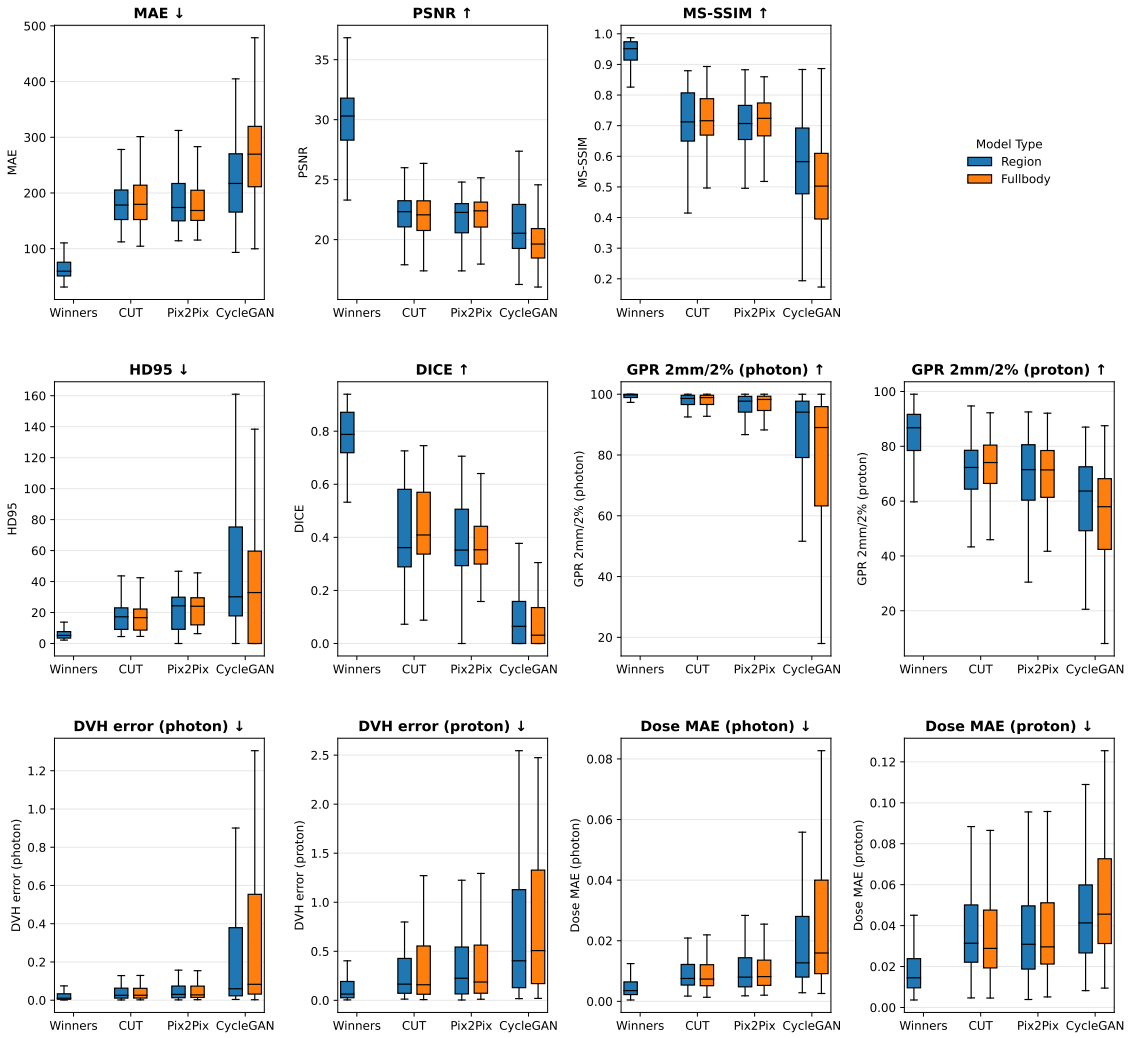


Figure 6.3: Experiment 1 Results Based on SynthRAD2025 Post-Challenge Evaluation. Experiment 1 results aggregated over three anatomical regions (abdomen, head-neck, thorax) based on SynthRAD2025 post-challenge evaluation on the private SynthRAD test set (223 volumes).

6.2 Experiment 2: Input Normalization Techniques (RQ2)

The second experiment investigates the impact of MRI intensity normalization and model-based domain adaptation on sCT quality for abdominal patients. This experiment isolates the preprocessing strategy as the primary variable, while additionally evaluating an architectural alternative to normalization-based harmonization. The following research question is addressed:

> *RQ2: How do different MRI intensity normalization techniques affect sCT quality, and can tissue-based normalization improve performance over global approaches in the presence of multi-center scanner heterogeneity?*

If normalization can reduce this distributional shift, the translation model should benefit from more consistent intensity – tissue correspondences across the training set.

6.2.1 Setup

Data. The experiment is restricted to the abdominal subset of the SynthRAD dataset, comprising 175 patients. Restricting the experiment to a single anatomical region eliminates confounding effects from anatomical heterogeneity and allows isolating the effect of normalization on scanner-induced intensity variability. Preprocessing follows Section 5.1, several normalization strategies are compared.

Normalization strategies. The following four MRI normalization strategies are compared, representing approaches of increasing sophistication to harmonizing MRI intensities (see Section 3.1 and Section 5.1 for detailed descriptions):

- *Baseline*: global clipping + rescaling
- *P99*: per-image percentile clipping + rescaling
- *Nyúl*: piecewise linear histogram standardization
- *N-Peaks*: tissue-specific peak alignment followed by piecewise linear histogram standardization

CT normalization remains unchanged across all conditions (HU clipping to $[-1024, 1200]$ followed by min-max scaling to $[0, 1]$).

Nyúl normalization. Nyúl normalization was implemented using the intensity-normalization package with the 2nd and 98th percentiles as boundary landmarks ($p_{\min}$, $p_{\max}$), chosen to exclude extreme outliers while retaining the full clinically relevant intensity range (see Section 3.1).

N-Peaks configuration. Three N-Peaks parameters are varied: LIC threshold $q \in \{0.3, 0.8\}$ (see Equation 3.5), number of N4 bias field correction passes (0, 1, or 2) (see Section 3.1.3), and peak estimation scope (global vs. center-specific), yielding $2 \times 3 \times 2 = 12$ configurations. Both scope settings map tissue peaks to a shared target scale; the difference lies in how target landmarks are estimated. The global setting pools peaks across all training patients, yielding target intensities that reflect a cross-center average. The center-specific setting estimates peaks separately per acquisition center, producing more precise per-center landmarks that account for systematic shifts in raw peak locations due to differences in field strength and pulse sequence. Both

are evaluated to assess whether this additional granularity translates into a measurable benefit. The full combinatorial space is not exhaustively evaluated in model training, as each configuration requires a full training run. Instead, a subset of four is selected. Additional considerations regarding parametrization and quality of results for N-Peaks are provided in the appendix (Figures A.8-A.11).

For the liver mask, center-specific configurations produce three clearly separated peaks corresponding to the three acquisition centers, while global configurations collapse these into a single shared peak. For the fat masks, no clear or consistent peaks are observable across any configuration, likely due to partial-volume effects, the diverse depot types in the combined mask, and scanner-specific contrast differences. As a result, fat does not serve as a reliable anchor in this dataset. The overlay in Figure A.5 further confirms that the overall distributions are highly similar across all twelve configurations, with meaningful differences confined to the liver peak region.

Based on this analysis, four configurations are selected for downstream model training: $q \in \{0.3, 0.8\}$ combined with both scope settings, each with a single N4 pass. Although the visual distributions show no strong evidence that N4 correction improved peak alignment, a single iteration is generally recommended in the literature as it reduces low-frequency field inhomogeneity without introducing substantial distortion. Since the four selected configurations produce closely comparable distributions, Table 6.4 and Figure 6.5 report only a single representative configuration (N-Peaks 0.8 LIC, 1x N4 Bias field correction, global scope) for conciseness.

Tissue intensity distributions. Figure 6.4 shows the resulting MRI intensity distributions across all normalization strategies. The progression from raw to increasingly sophisticated normalization is visible across all three tissue masks. Baseline normalization compresses the body contour distributions by removing potential outliers, effectively stretching the remaining intensity range. P99 intensifies this effect, and notably, three peaks begin to emerge in the liver region, likely corresponding to the three acquisition centers. Nyúl normalization attempts to map all patients onto a common scale, merging these center-specific peaks toward a single shared distribution. N-Peaks then produces a clearly defined liver peak, as expected given that liver serves as the alignment target. The fat mask, however, does not exhibit a comparably well-defined peak under any configuration. Comparing the global and center-specific N-Peaks settings confirms this asymmetry: for the liver, the center-specific approach yields three distinct peaks rather than one, whereas the fat distributions remain largely unchanged between the two settings.

Separate first layer. In addition to the preprocessing-based normalization strategies, the *separate first layer* approach described in Section 5.2.1 is evaluated. While this architecture-level strategy could in principle be combined with preprocessing normalization, it is evaluated independently here to isolate its effect relative to the input-level harmonization methods.

Training and evaluation. To assess whether the effect of normalization is consistent across translation paradigms, all normalization strategies are evaluated across the three image-to-image translation frameworks: Pix2Pix, CUT, and CycleGAN. The same baseline generator architecture and training configuration (learning rate 0.0002, batch size 1, number of epochs 50) as in Experiment 1 are used for each respective framework. The baseline normalization models are reused from Experiment 1 and not trained again. Postprocessing and evaluation follow the same protocol as Experiment 1. A limited dosimetric evaluation is also performed on three test-set patients.

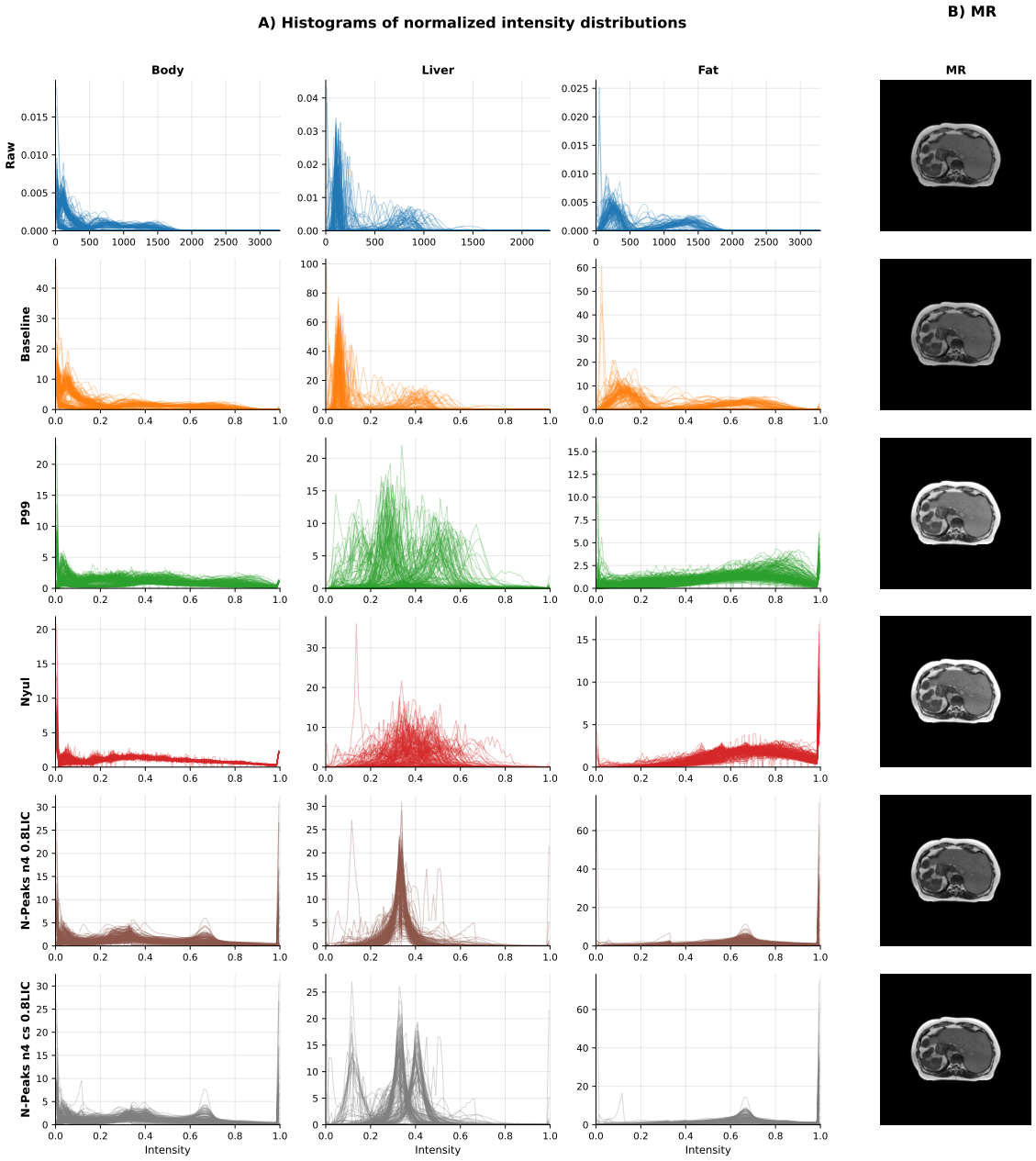


Figure 6.4: TISSUE-SPECIFIC MRI INTENSITY DISTRIBUTIONS ACROSS NORMALISATION STRATEGIES. Tissue-specific MRI intensity distributions for the abdominal training set (122 volumes) across all normalisation strategies. Per-patient kernel density estimates are shown for the body contour, liver, and fat masks (A), alongside a representative axial MR slice for each method (B). Rows correspond to raw intensities, Baseline, P99, Nyúl, and two N-Peaks configurations varying peak estimation scope (global vs. center-specific).

6.2.2 Results

The results of Experiment 2 are summarized in Table 6.4 and for easier readability of overall trends in Figure 6.5. The relative ordering of translation frameworks observed in Experiment 1 is reproduced: CUT and Pix2Pix perform comparably and substantially outperform CycleGAN across all normalization conditions. The Wilcoxon signed-rank test is applied to compare each normalization strategy against the Baseline within each framework.

N-Peaks achieves the best or near-best performance across all three frameworks. For CUT, it yields the lowest MAE (113.56 ± 25.27) and MSE, with a statistically significant improvement over Baseline ($p < 0.01$). For Pix2Pix, N-Peaks (106.85 ± 28.23) performs on par with Nyúl (106.82 ± 29.24), both significantly improving over Baseline ($p < 0.01$). For CycleGAN, N-Peaks again produces the lowest MAE (149.97 ± 30.28, $p < 0.01$), though all CycleGAN configurations remain substantially worse than the paired frameworks.

The remaining normalization strategies exhibit framework-dependent effects. Nyúl significantly improves Pix2Pix (lowest MAE and highest SSIM, $p < 0.01$) but significantly degrades CUT performance relative to Baseline ($p < 0.01$). P99 normalization improves CycleGAN ($p = 0.01$) but degrades Pix2Pix ($p = 0.01$), indicating that a simple global rescaling interacts differently with each framework's training dynamics. These framework-normalization interactions underscore that no universal preprocessing strategy exists; the choice of normalization must be considered jointly with the translation framework.

The separate first layer approach consistently underperforms across all three frameworks. Rather than improving performance, it substantially degrades results, with CUT and CycleGAN exhibiting particularly high variance, indicating training instability. In fact, the comparison of results stratified by treatment center as provided in Appendix A.5 shows, that this is mainly caused by center C for which only 13 training and 3 testing samples are available.

Additionally, a dosimetric analysis is performed on a subset of three abdominal test cases to assess the clinical relevance of the generated sCT images beyond voxel-wise similarity metrics. Using a treatment planning pipeline based on matRad, dose distributions are calculated on both the reference CT and the corresponding sCT, and DVH-derived delta metrics are computed to quantify systematic deviations. Friedman tests across all models and normalization strategies yield p-values above the significance threshold of $p = 0.05$, indicating no statistically significant differences between normalization strategies — a result likely attributable to the small sample size. For Pix2Pix and CUT, the majority of DVH indicator differences remain within the $\pm 2\%$ clinical acceptability threshold, suggesting that the choice of normalization strategy has limited dosimetric impact for these models. The full results of this analysis are provided in Appendix A.7.

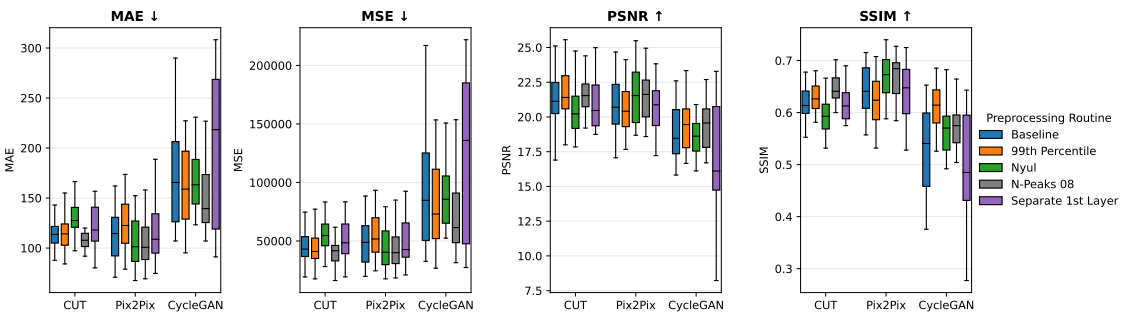


Figure 6.5: Experiment 2 Results on Abdominal Data based on own image similarity evaluation, test set (27 volumes).

Table 6.4: EXPERIMENT 2 RESULTS ON ABDOMINAL DATA BASED ON OWN IMAGE SIMILARITY EVALUATION (MEAN ± STD), TEST SET (27 VOLUMES). Wilcoxon signed-rank test against baseline preprocessing routine of the respective framework; p values are not reported if symmetry assumption seems violated (separate 1st layer). The best-performing model within each framework is highlighted in bold; the overall best model is additionally marked in italic.

Framework	Preprocessing Routine	MAE ↓	Wilcoxon p (MAE)	MSE ↓	PSNR ↑	SSIM ↑
CUT	Baseline	119.20 ± 29.66	–	48′703 ± 29′054	21.38 ± 1.77	0.620 ± 0.031
CUT	99th Percentile	117.37 ± 27.29	0.52	46′183 ± 25′365	21.63 ± 1.77	0.627 ± 0.027
CUT	Nyúl	132.47 ± 24.92	0.00	59′346 ± 23′441	20.61 ± 1.78	0.593 ± 0.034
CUT	N-Peaks 08	**113.56 ± 25.27**	0.00	***44′672 ± 23′800***	***21.71 ± 1.65***	**0.646 ± 0.028**
CUT	Separate 1st Layer	201.50 ± 241.4	–	136′323 ± 256′716	19.73 ± 4.37	0.578 ± 0.121
Pix2Pix	Baseline	116.94 ± 35.40	–	53′867 ± 30′646	20.84 ± 1.95	0.644 ± 0.049
Pix2Pix	99th Percentile	126.33 ± 31.28	0.01	57′578 ± 26′235	20.47 ± 1.80	0.622 ± 0.047
Pix2Pix	Nyúl	***106.82 ± 29.24***	0.00	45′535 ± 24′270	**21.60 ± 1.87**	***0.669 ± 0.044***
Pix2Pix	N-Peaks 08	106.85 ± 28.23	0.00	**45′462 ± 23′803**	21.51 ± 1.73	0.666 ± 0.043
Pix2Pix	Separate 1st Layer	117.22 ± 34.97	–	53′613 ± 28′901	20.77 ± 1.84	0.640 ± 0.055
CycleGAN	Baseline	174.47 ± 54.34	–	94′561 ± 51′400	18.84 ± 2.10	0.528 ± 0.082
CycleGAN	99th Percentile	159.64 ± 40.22	0.01	80′243 ± 34′677	19.36 ± 1.90	**0.608 ± 0.045**
CycleGAN	Nyúl	168.81 ± 34.36	0.33	89′789 ± 29′866	18.47 ± 1.44	0.565 ± 0.042
CycleGAN	N-Peaks 08	**149.97 ± 30.28**	0.00	**70′867 ± 28′632**	**19.44 ± 1.68**	0.571 ± 0.037
CycleGAN	Separate 1st Layer	221.77 ± 145.3	–	143′566 ± 150′188	17.29 ± 3.56	0.504 ± 0.100

6.3 Experiment 3: Alternative Generator Networks (RQ3)

The third experiment investigates the impact of different generator architectures and training regimes within a fixed translation framework. Specifically, the following research question is addressed:

> *RQ3: How does the choice of generator architecture and training regime influence sCT quality within the Pix2Pix framework, and how does adversarial training compare to a generator-only setting?*

The motivation is to assess whether architectural advances beyond the standard UNet generator can improve performance, and whether adversarial training is essential when using more expressive backbone networks such as Swin-based transformers. In particular, the experiment examines convolutional (UNet and ResNet) versus transformer-based (SwinV2-UNet) generators, and compares cGAN training to a pure regression (generator-only) setting for transformer-based generators.

6.3.1 Setup

Data. To reduce anatomical variability and isolate architectural effects, this experiment is restricted to the *abdominal* subset of the SynthRAD dataset. Preprocessing follows the baseline configuration in Section 5.1 without modification. The abdominal Pix2Pix baseline model from Experiment 1 (UNet generator) is reused and not retrained.

Model Variants. All architectural modifications are implemented within the Pix2Pix framework. In addition to the abdominal Pix2Pix UNet baseline, the following generator variants are evaluated:

- **Pix2Pix + ResNet generator:** Replacement of the UNet backbone with a ResNet-based encoder-decoder architecture.
- **Pix2Pix + SwinV2-UNet (frozen encoder):** SwinV2-UNet generator where all encoder layers are frozen except for the first linear embedding layer, which is adapted due to input channel mismatch.
- **Pix2Pix + SwinV2-UNet (fully trainable):** SwinV2-UNet generator with all layers being trainable.

To disentangle architectural effects from adversarial training, a *generator-only* setting is additionally evaluated (all other training components such as data loading, augmentation, batching remain unchanged):

- **Generator-only + SwinV2-UNet (frozen encoder):** Same architecture as above, trained using only the L_1 reconstruction loss without discriminator.
- **Generator-only + SwinV2-UNet (fully trainable):** Fully trainable SwinV2-UNet optimized solely with the L_1 loss.

Table 6.5 summarizes the key hyperparameters and model sizes for all evaluated variants in Experiment 3. For the SwinV2-based generators, the initial baseline learning rate of 2×10^{-4} led to unstable training and non-convergence; therefore, it was reduced to 2×10^{-5}. Due to the slower observed convergence behavior (independently of freezing the encoder-part of the network) of the transformer-based models, the number of training epochs was increased from 50 to 100 to ensure sufficient optimization time.

Table 6.5: Experiment 3 Overview of Key Hyperparameters and Model Sizes.

Framework	Generator Backbone	Learning Rate	# Epochs	# Params (G)	# Trainable Params (G)
Pix2Pix	UNet-256 (baseline)	0.0002	50	54.41 M	54.41 M
Pix2Pix	ResNet-9	0.0002	50	11.37 M	11.37 M
Pix2Pix	SwinV2-UNet (fully trainable)	0.00002	100	34.35 M	34.35 M
Pix2Pix	SwinV2-UNet (frozen encoder)	0.00002	100	34.35 M	6.78 M
Generator-only	SwinV2-UNet (fully trainable)	0.00002	100	34.35 M	34.35 M
Generator-only	SwinV2-UNet (frozen encoder)	0.00002	100	34.35 M	6.78 M

Postprocessing and Evaluation. Postprocessing is identical to previous experiments. Evaluation is restricted to our own image similarity metrics (see Section 3.5). No SynthRAD2025 post-challenge submission (due to restricting to abdominal region) and no dosimetric evaluation (due to computational constraints) are performed for this experiment.

6.3.2 Results

The quantitative results of Experiment 3 are summarized in Table 6.6 and visualized for easier readability of overall trends in Figure 6.6.

Within the Pix2Pix framework, replacing the UNet-256 baseline with a ResNet-9 generator leads to consistent improvements across all four image similarity metrics. ResNet-9 achieves the lowest MAE and MSE as well as the highest PSNR and SSIM, indicating both reduced voxel-wise error and improved perceptual similarity. The fully trainable SwinV2-UNet also improves over the UNet baseline in terms of MAE, MSE, and PSNR, but remains inferior to ResNet-9. In

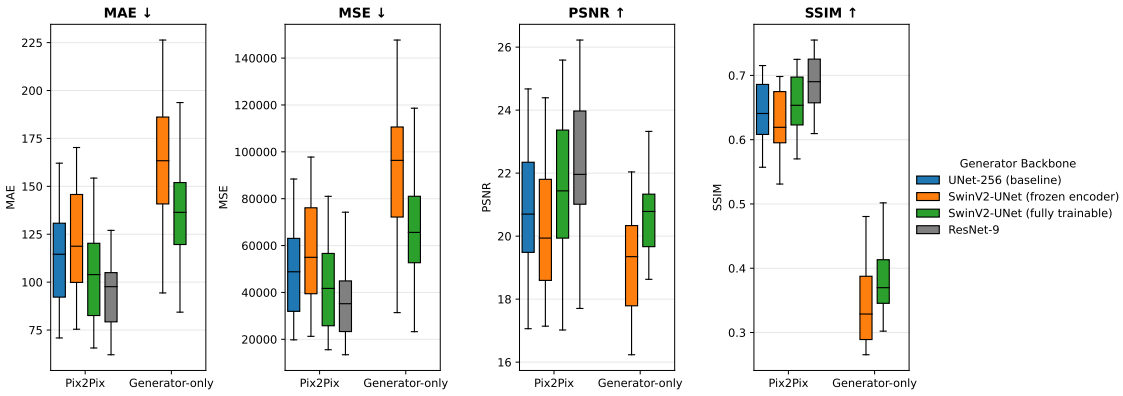


Figure 6.6: EXPERIMENT 3 RESULTS ON ABDOMINAL DATA. Evaluation based on own image similarity implementation, test set (27 volumes).

contrast, freezing the SwinV2 encoder results in a performance degradation compared to both the fully trainable SwinV2-UNet and the UNet baseline, suggesting that end-to-end optimization is essential for transformer-based generators in this setting (randomly initialized first layer).

The generator-only configurations exhibit substantially inferior performance compared to all adversarially trained Pix2Pix models. Both SwinV2-UNet variants trained solely with the L_1 loss show markedly increased MAE and MSE and pronounced reductions in SSIM. Although the fully trainable generator-only model partially recovers PSNR relative to the frozen variant, it remains clearly below the corresponding Pix2Pix configuration. This indicates that adversarial training contributes significantly to structural realism and perceptual consistency beyond pure voxel-wise regression.

Table 6.6: EXPERIMENT 3 RESULTS ON ABDOMINAL DATA BASED ON OWN IMAGE SIMILARITY EVALUATION. Experiment 3 results on abdominal data based on own image similarity evaluation (mean ± std), test set (27 volumes). The best-performing model within each framework is highlighted in bold; the overall best model is additionally marked in italic.

Framework	Generator Backbone	MAE ↓	MSE ↓	PSNR ↑	SSIM ↑
Pix2Pix	UNet-256 (baseline)	116.94 ± 35.40	53′867 ± 30′646	20.84 ± 1.95	0.644 ± 0.049
Pix2Pix	SwinV2-UNet (frozen encoder)	125.06 ± 36.44	60′394 ± 31′858	20.35 ± 2.06	0.627 ± 0.050
Pix2Pix	SwinV2-UNet (fully trainable)	108.24 ± 35.93	46′613 ± 30′811	21.55 ± 2.14	0.655 ± 0.048
Pix2Pix	ResNet-9	***98.51 ± 31.9***	***39′502 ± 27′219***	***22.27 ± 2.13***	***0.689 ± 0.047***
Generator-only	SwinV2-UNet (frozen encoder)	164.91 ± 32.27	92′217 ± 28′674	19.21 ± 1.54	0.347 ± 0.065
Generator-only	SwinV2-UNet (fully trainable)	**135.34 ± 24.31**	**66′205 ± 22′444**	**20.67 ± 1.29**	**0.382 ± 0.055**

Overall, the results indicate that architectural choice has an impact on sCT quality within the Pix2Pix framework. In the investigated abdominal setting, the convolutional ResNet-9 backbone outperforms both the UNet baseline and the transformer-based SwinV2-UNet variants. Furthermore, removing the adversarial objective leads to a substantial performance drop, highlighting the importance of cGAN training for realistic MR-to-CT translation in this configuration. With respect to RQ3, the findings suggest that both generator architecture and training regime critically influence performance, and that architectural expressiveness alone does not compensate for the absence of adversarial supervision.

Chapter 7

Discussion

This chapter reflects on the presented work from perspectives that a results-oriented chapter cannot fully address. It examines overarching methodological choices, including preprocessing strategies, data alignment, and architectural decisions, in terms of their broader implications for clinical translation and future research. Where certain limitations or open questions are best understood in the context of a specific experiment, they are discussed accordingly.

Derivation of Body Contours and Organ Delineations. For training, validation and image similarity-based evaluation, CT-based thresholding was used to derive body contours, as the masks provided by the SynthRAD organizers were comparatively coarse and occasionally included non-anatomical regions. Under the assumption that clinically deployed systems would rely on accurate contouring (e.g. from dedicated planning workflows), this tighter masking strategy likely improved the precision of voxel-wise supervision during training. However, for official SynthRAD submissions, evaluation relied on the organizer-defined masks rather than the refined CT-based contours used internally. This discrepancy may have contributed to performance differences between internal and official evaluations. It should further be noted that CT-based thresholding is not available during inference in an MRI-only workflow, and MR-based thresholding is considerably more challenging due to scanner- and protocol-dependent intensity variability (see Section 4.4), limiting the direct transferability of this masking strategy to a purely MR-driven deployment scenario.

For dosimetric evaluation, anatomical structures were segmented on CT volumes using TotalSegmentator, as skin delineation is currently implemented only for CT input. In a clinically realistic MRI-only setting, however, target volumes and organs at risk would need to be delineated directly on MR images or sCT. Consequently, the adopted approach constitutes a practical surrogate for experimental purposes.

Aligned vs. Unaligned Data. An important methodological aspect concerns the use of aligned MRI-CT pairs. Pix2Pix explicitly exploits paired and spatially aligned data through its conditional adversarial formulation and L_1 reconstruction loss. In contrast, CycleGAN and CUT do not require paired supervision and, in principle, do not leverage the exact voxel-wise correspondence available in the dataset. Despite this, CUT with a ResNet backbone achieves performance comparable to Pix2Pix-UNet, suggesting that the architectural choice may be more influential than strict exploitation of paired alignment in this setting. Moreover, the observation that Pix2Pix-ResNet outperforms Pix2Pix-UNet, while CUT-ResNet performs similarly to Pix2Pix-UNet without using paired supervision, indicates that residual architectures may be well suited for MR-to-CT translation.

Training Effort vs. Performance (CUT). CUT was trained substantially longer per epoch due to its contrastive learning objective and patch-based feature sampling (roughly 20min per epoch for Pix2Pix compared to 2h per epoch for CUT). Despite this increased computational effort, its performance remains largely comparable to Pix2Pix and does not consistently outperform it. While CUT demonstrates competitive perceptual quality and favorable dosimetric stability, the absence of a clear quantitative advantage raises questions about its training efficiency in this specific MR-to-CT setting. In particular, the additional training complexity does not translate into a decisive improvement in MAE or PSNR. However, in a clinical setting after model deployment inference time is relevant, indicating this aspect is only relevant if computational resources for model training are a bottleneck. Nevertheless, comparable performance of CUT and Pix2Pix remains notable since CUT does not exploit the availability of aligned data.

Output Activation Function. Following the original Pix2Pix, CycleGAN, and CUT implementations, a hyperbolic tangent (`tanh`) activation function was applied to the generator output, mapping predictions to the range $[-1, 1]$. To reconcile this with the normalized $[0, 1]$ input representation used during training, a linear rescaling is applied (both in the original implementations during training and in our pipeline at inference). While this mismatch is technically resolved through such rescaling, the approach carries inherent limitations that warrant discussion. Most notably, the `tanh` function is asymptotic: it approaches its bounds only in the limit, meaning that the extreme values of the clipped HU range employed for training (-1024 HU and 1200 HU) are in practice unattainable. The network can only predict values *approaching* these extremes, which may introduce systematic errors for very dense structures (e.g. bones) or very low-density regions (e.g. air-filled cavities). Beyond this numerical concern, the reliance on `tanh` with post-hoc linear rescaling is rather unintuitive. For sCT generation, a more principled alternative such as a linear output activation, or a sigmoid activation would better reflect the regression nature of the problem.

7.1 Experiment 1

Qualitative Results. The qualitative behavior of the different frameworks has also been investigated using randomly chosen examples of axial and coronal slices from the same volumes. While axial views allow direct comparison in the training plane, the reconstructed coronal slices provide insight into cross-slice consistency and three-dimensional coherence of the generated sCT volumes.

The axial examples illustrated in Figure 7.1 exhibit attributes that are interesting to compare qualitatively. The yellow region marks air, which is reproduced with varying fidelity across the generated outputs; some align more closely with the MR appearance than others. Note that the ground truth CT was acquired at a different time point, so MR-CT differences are expected. The blue region highlights the ribs, which are largely absent in the CUT and CycleGAN outputs and inconsistently represented in the Pix2Pix results, with variation in rib number and position compared to the ground truth CT. The CycleGAN outputs appear flipped upside-down with the spinal area (red arrow) on the wrong side for this specific slice.

Figure 7.2 shows coronal reconstructions of the same volumes as the axial examples. Additional interesting qualitative behaviors can be assessed from a cross-slice consistency perspective. Pix2Pix outputs show cross-slice inconsistencies that produce noticeable blurring and structural irregularities, whereas CUT appears more three-dimensionally consistent overall, though blurring remains in regions with sharp intensity transitions (e.g., lung borders). The yellow region highlights the inferior portion of the right lung, where different types of blur are evident across the model outputs (CycleGAN appears to have learned this portion most consistently). The blue

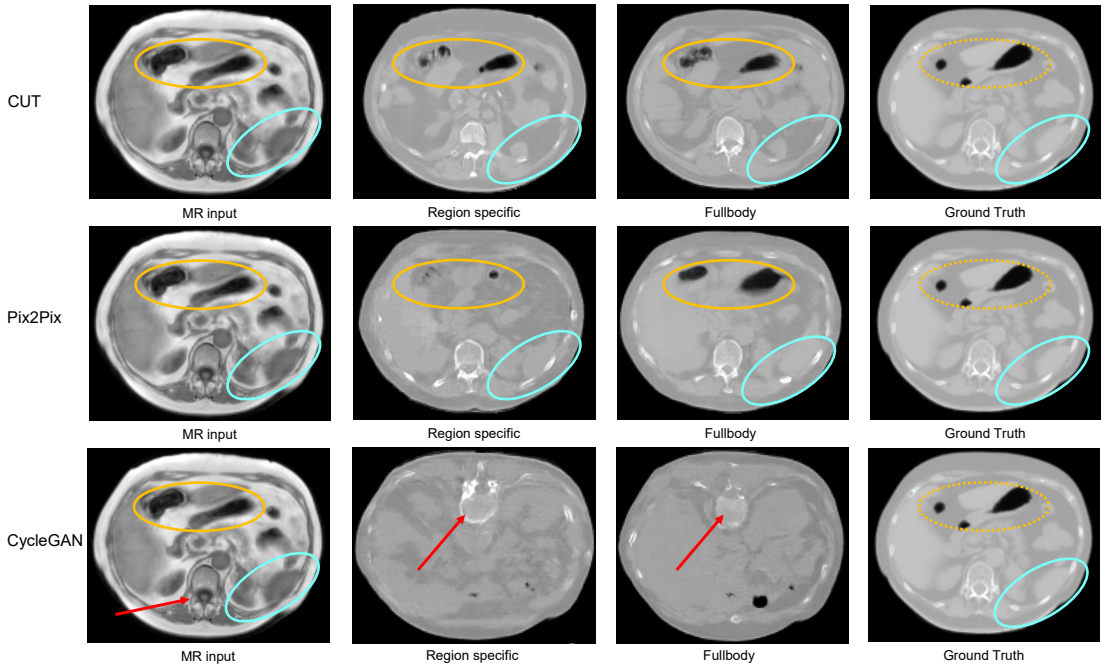


Figure 7.1: QUALITATIVE COMPARISON OF MODEL OUTPUTS ON AN ABDOMINAL AXIAL SLICE. Example abdominal axial slice showing the MR input, model outputs (full body and region specific), and the corresponding ground truth CT (same random slice for all models).

region marks the ribs, which are largely absent in the CUT and region specific Pix2Pix outputs, while the full body Pix2Pix output appears to overestimate bone content in this area. CycleGAN exhibited an artificial horizontal banding pattern in the coronal view, which arose because adjacent axial slices were nearly identical, causing the model to reproduce slice-wise similarities as structured lines after coronal reconstruction (red box).

Generalization vs. Region-Specific Specialization. Experiment 1 was designed to investigate the trade-off between model generalization and region-specific specialization. Overall, the results suggest that joint training across all five anatomical regions did *not* lead to a substantial performance gain compared to region-specific models. These findings indicate that a single model can learn a reasonably robust cross-region representation without dramatic loss in performance, at least for Pix2Pix and CUT. However, modest but consistent improvements for region-specific models-particularly for Pix2Pix-suggest that specialization still provides benefits when the goal is to optimize voxel-wise accuracy. Our findings align with the feasibility claims in Hsu et al. (2025). However, it is important to note that performance remained comparable at best, and slightly inferior in several metrics.

Region-Dependent Resampling and Scale Variation. An additional factor that may have complicated cross-region learning in the fullbody setting concerns the region-dependent resampling strategy described in Section 5.1.2. Head-and-neck and brain volumes are cropped directly to 256×256 without downsampling, while thorax, abdomen, and pelvis volumes are first standardized to 512×512 and then subsampled by a factor of two. As a consequence, anatomical

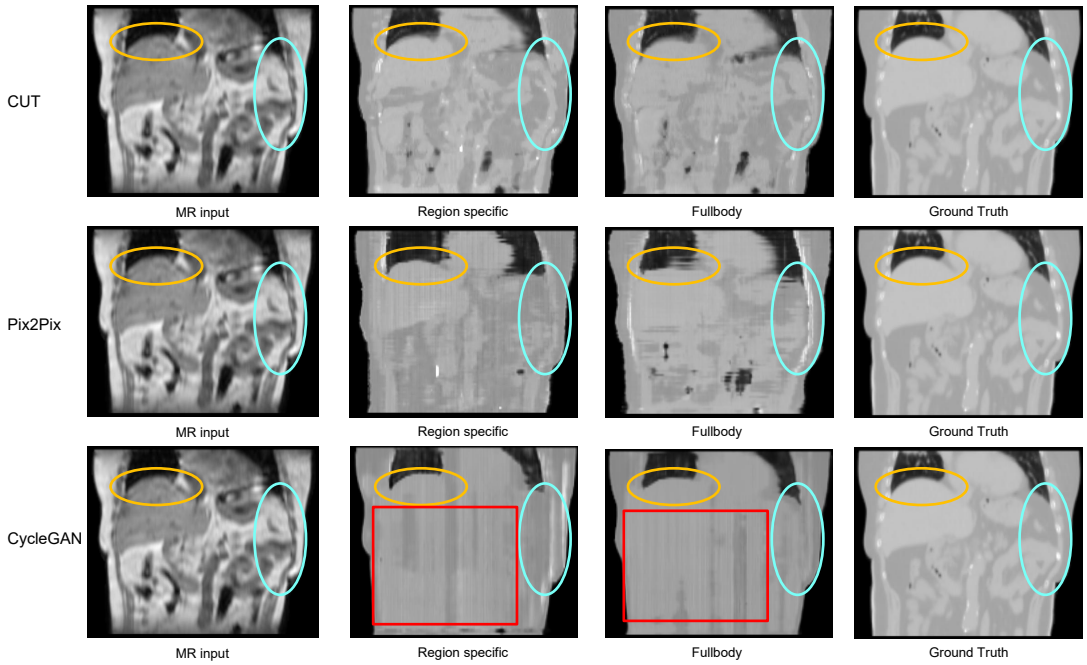


Figure 7.2: QUALITATIVE COMPARISON OF MODEL OUTPUTS ON AN ABDOMINAL AXIAL SLICE. Example abdominal coronal slice showing the MR input, model outputs (fullbody and region-specific), and the corresponding ground truth CT (same random slice for all models).

structures appear at approximately twice the spatial resolution per axis in head-and-neck and brain slices compared to the remaining body regions. When all regions are pooled into a single model, the generator must therefore learn to translate features that vary substantially in apparent scale across the training distribution, which may increase the difficulty of learning a coherent cross-region representation. This discrepancy arises from a deliberate trade-off: retaining the finer pixel spacing of smaller fields of view preserves more anatomical detail on the input image, whereas enforcing a uniform physical resolution across all regions would require upsampling the larger body scans or downsampling the smaller ones, in either case discarding information or introducing interpolation artifacts. Further investigations of options and implications are discussed in Appendix Section A.1.

Internal Evaluation vs. Official SynthRAD Evaluation. A noticeable discrepancy was observed between our internal image similarity evaluation and the official SynthRAD2025 post-challenge results. First, the official evaluation aggregated only three anatomical regions (abdomen, head-and-neck, thorax), whereas our internal analysis included five regions. Second, the evaluation pipelines differ in preprocessing details and metric implementation (e.g. MS-SSIM instead of aggregated 2D SSIM). Moreover, the challenge evaluation operates on organizer-defined masks and resampling conventions, while our internal evaluation used our own CT-derived body masks and preprocessing pipeline.

Most strikingly, Pix2Pix exhibits a considerable increase in MAE when moving from our internal evaluation to the official SynthRAD evaluation. Possible contributing factors include differences in anatomical coverage, masking, dataset, or intensity clipping. As shown in Section A.4, the majority of the discrepancy for Pix2Pix-Fullbody can be attributed to anatomical coverage

(only abdomen, head-neck and thorax) as well as differences in masking. The absence of a similar magnitude shift for CUT suggests framework-dependent sensitivity to these evaluation conditions.

Beyond architectural differences, the computational cost is strongly shaped by the training setup. A fullbody model is trained on slices pooled from all five anatomical regions, which increases the amount of data seen per epoch and typically leads to longer epochs and higher overall runtime until convergence. However, this fullbody model cost should be interpreted in context: the relevant baseline is not one region-specific model, but training and maintaining five separate region-specific models. From that perspective, the comparison becomes a trade-off between (i) one more expensive training run with a single set of weights versus (ii) multiple cheaper runs that together may match or exceed the total training effort, depending on the number of regions used, stopping times, and whether training can be parallelized. Training the five region-specific models for Experiment 1 in our setting has required approximately the same resource-time as training the fullbody model, but parallelizing has reduced the time from initialization to convergence.

The observed performance differences remain small —- fullbody training is at best comparable to region-specific training across image similarity and dosimetric metrics. This indicates that pooling all regions into one model does not consistently yield a measurable accuracy advantage, so the practical choice is largely driven by efficiency and deployment considerations: a joint model can simplify inference and maintenance, while region-specific models may be preferable when the highest region-wise performance is desired and training resources allow multiple specialized runs.

Failure Mode of CycleGAN in the Abdominal Setting. A closer inspection of the abdominal CycleGAN outputs reveals a characteristic failure pattern that helps explain its comparatively poor quantitative performance. Instead of learning a consistent, anatomy-dependent MR-to-CT mapping, the model frequently generates slices resembling an *average abdominal CT slice located near the center of the volume*, largely independent of the actual input anatomy. This behavior suggests that the cycle-consistency constraint may have over-constrained the optimization: rather than encouraging a faithful conditional mapping, it potentially promotes a solution that minimizes the bidirectional reconstruction error by collapsing towards a prototypical appearance in CT space. In other words, the model appears to prioritize invertibility of the mapping over accurate, slice-specific CT synthesis. This interpretation is further supported by the observation that the generated average slice is occasionally flipped upside-down, consistent with the random flipping augmentations applied during training. This default behavior of the CycleGAN implementation seems to be highly destructive in the presence of non-symmetric, paired data. Such symmetry-consistent artifacts indicate that the generator does not anchor its prediction to the concrete spatial configuration of the input slice, but instead converges to a mode that satisfies cycle consistency while neglecting anatomical correspondence.

Performance Gap to Challenge Winners. All investigated configurations remained substantially below the official challenge winners across image similarity, geometric, and dosimetric metrics. The winners achieved markedly lower MAE and higher PSNR and MS-SSIM, as well as superior DICE and HD95 values. Two methodological differences likely contributed to this gap. First, top-performing challenge methods typically employ 3D architectures, which inherently model inter-slice context and reduce cross-slice inconsistencies observed in our 2D reconstructions (see Figure 7.2 and Figure 7.4). These approaches, however, require significantly more computational resources than the approaches analyzed in this experiment. Second, many leading approaches incorporate perceptual or feature-based losses in addition to adversarial and pixel-wise objectives, which can improve structural realism and boundary sharpness beyond L_1-based optimization.

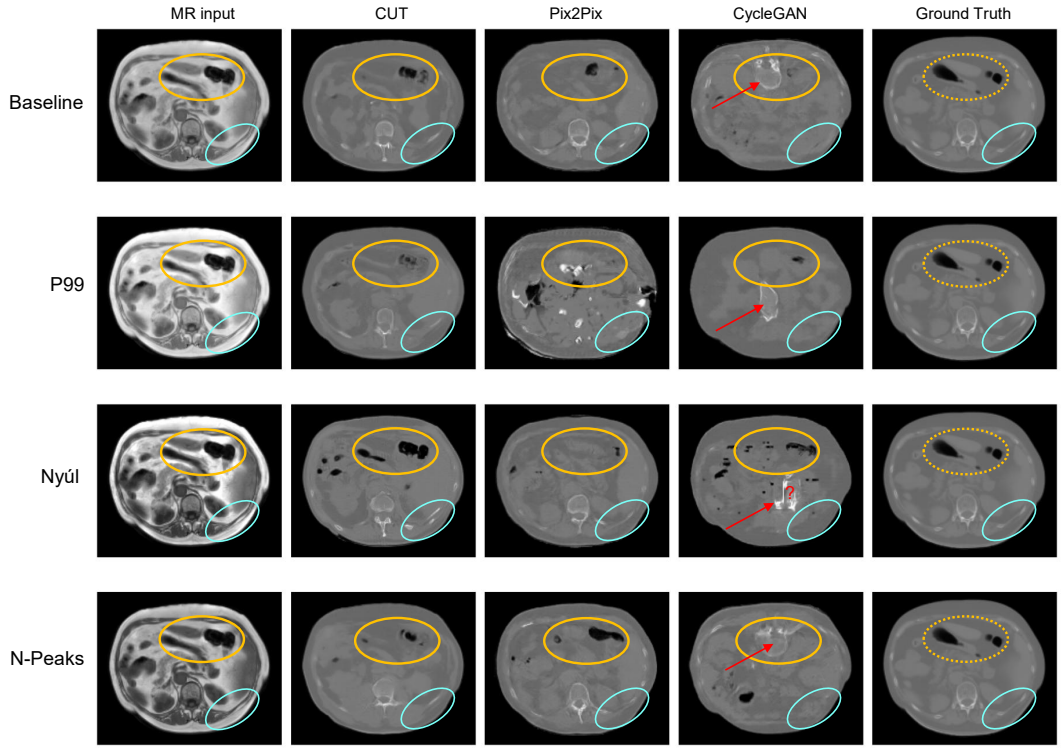


Figure 7.3: QUALITATIVE COMPARISON OF MODEL OUTPUTS ACROSS NORMALIZATION STRATEGIES ON AN ABDOMINAL AXIAL SLICE. Example abdominal axial slice showing the MR input (left column), outputs from CUT, Pix2Pix, and CycleGAN (middle columns), and the corresponding ground truth CT (right column) for each of the four normalization strategies (same patient and slice across all rows).

7.2 Experiment 2

Qualitative Results. The qualitative behavior of the different frameworks has also been investigated using randomly chosen examples of axial slices as displayed in Figure 7.3. The yellow region marks air-filled structures in the upper abdomen, which are reproduced with varying fidelity across frameworks; note that the ground truth CT was acquired at a different time point, so MR-CT differences in bowel gas distribution are expected. The cyan region highlights the lateral soft tissue and musculature area. The P99 MR input appears noticeably brighter overall compared to the other normalization strategies, reflecting its global intensity rescaling. The N-Peaks MR input, by contrast, shows reduced intensity in subcutaneous fat regions, consistent with its tissue-specific peak alignment. CycleGAN outputs exhibit spatial distortions across all normalization strategies, with the spinal area displaced toward the anterior side (red arrows), consistent with the flipping artifacts observed in Experiment 1. CUT and Pix2Pix produce outputs that more closely follow the anatomical layout of the MR input, though Pix2Pix renders slightly sharper tissue boundaries.

Limited overall impact of normalization. The relatively small performance differences between normalization strategies indicate that modern GAN-based translation frameworks are to

some degree robust to the form of MRI intensity preprocessing, at least when evaluated by global image similarity metrics. The translation model appears to accommodate residual inter-patient and inter-center intensity variability without requiring precise distributional alignment as a precondition. This robustness may be partly attributable to the adversarial training objective, which encourages the generator to produce plausible CT-like outputs regardless of the specific form of the input distribution.

Normalization effects are framework-dependent. Among the tested strategies, no method consistently outperformed the others across all frameworks and metrics. This framework-dependent pattern is particularly evident for Nyúl normalization: its piecewise linear histogram alignment appears to benefit Pix2Pix's pixel-wise paired supervision while interfering with CUT's contrastive patch-based objective, which may be more sensitive to global intensity reshaping that alters spatial contrast patterns. N-Peaks, as a tissue-anchored refinement of the Nyúl principle, shows more consistent behaviour across frameworks, though its advantage is constrained by the unreliability of fat tissue as a normalization anchor in the abdominal region, as discussed above.

An interesting divergence was observed between CUT and Pix2Pix in their response to Nyúl normalization. For Pix2Pix, Nyúl yields a clear improvement over Baseline; for CUT, it leads to a notable degradation. This suggests that the two frameworks differ in how they exploit the statistical structure of the input distribution, and that the piecewise linear histogram alignment imposed by Nyúl may interact differently with paired and unpaired training objectives. Specifically, CUT's contrastive patch-based objective may be more sensitive to global intensity reshaping that alters spatial contrast patterns, whereas Pix2Pix's pixel-wise L_1 supervision appears to benefit from the more consistent intensity scale. This framework-normalization interaction warrants further investigation, particularly in multi-center settings where the most appropriate normalization strategy may depend on the chosen training paradigm.

Failure of the separate first layer. The separate first layer approach consistently underperforms across all frameworks and exhibits substantially higher variance, indicating training instability rather than a principled adaptation to center-specific characteristics. A major challenge for this center-specific adaptation strategy was the limited data availability per institution, particularly for center C. The abdominal dataset comprised only 13 training and 3 test cases for this center, which substantially restricted the models' ability to learn robust and generalizable feature representations. In several configurations, especially for CUT and CycleGAN, the models failed to extract meaningful features for center C, as reflected by the markedly degraded image similarity metrics and high variance. In contrast, performance on centers A and B was more stable and consistent with the baseline results. A detailed comparison of results stratified by treatment center is provided in Appendix A.5.

7.3 Experiment 3

Qualitative Results. The model outputs were also evaluated qualitatively, the results are shown in Figure 7.4. The models were trained on axial slices; therefore, the coronal views are reconstructed from axial predictions. The outputs show cross-slice inconsistencies that produce noticeable blurring and structural irregularities for all generators. The region highlighted in yellow marks air, which is fairly aligned with the input in the case of ResNet but not well predicted by UNet and SwinV2-UNet. The blue region marks the ribs, which are largely absent in the baseline and SwinV2-UNet and fairly represented with ResNet. The pink region highlights the inferior portion of the right lung, which seems quite well represented with ResNet with details missing

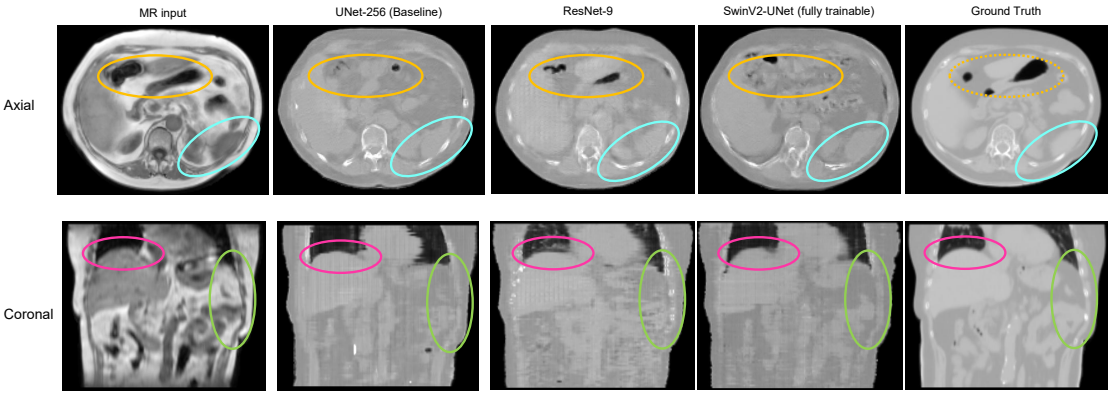


Figure 7.4: COMPARISON OF PIX2PIX MODELS WITH DIFFERENT GENERATOR NETWORKS ON ABDOMINAL DATA. Example abdominal axial and coronal slices showing the MR input, model outputs (Pix2Pix models trained with different generator networks on abdominal data), and the corresponding ground truth CT.

for UNet and SwinV2-UNet. Also for the ribs in the coronal view (green area), ResNet produces the most accurate representation. Generator-only results omitted due to poor results quality.

Convergence Behavior. The training time per epoch is comparable across UNet, ResNet, and SwinV2-UNet models (with the frozen Swin variant being slightly faster due to fewer trainable parameters). However, the Swin-based models require lower learning rates and more epochs to reach convergence, resulting in a higher overall training time. This slower convergence behavior may reflect the higher optimization complexity of transformer-based generators in a relatively limited-data, 2D slice-wise regime.

Frozen vs. Fully Trainable SwinV2-UNet. A particularly noteworthy observation is that the SwinV2-UNet with frozen encoder converges more slowly and achieves worse performance than the fully trainable variant, despite using the same reduced learning rate. One possible explanation is that adapting only the initial patch embedding layer while keeping deeper transformer blocks fixed constrained the model's ability to align pretrained feature representations with the MR-to-CT domain shift. In this configuration, the embedding layer may have needed to compensate for a mismatch between grayscale MR inputs and pretrained ImageNet features without support from jointly adapting deeper layers. At the same time, increasing the learning rate to accelerate embedding adaptation was not feasible, as it destabilized training of the randomly initialized decoder. This trade-off may explain the comparatively poor learning dynamics of the frozen configuration. Overall, end-to-end optimization appears essential for transformer-based generators in this application.

Resolving the Input Mismatch for Encoder Freezing. A conceptually cleaner approach to encoder freezing would be to replicate the single-channel grayscale MR input across all three input channels before passing it to the pretrained SwinV2 encoder, rather than adapting the patch embedding layer to accept single-channel input. Under this scheme, the encoder receives a three-channel tensor whose channels are identical, which is consistent with the RGB inputs seen during ImageNet pretraining and therefore allows the full encoder to be frozen without any modifica-

tion to its weights. The key advantage of this strategy is that the pretrained feature hierarchy remains entirely intact: no part of the encoder needs to be retrained or adapted, and the learning rate conflict between the patch embedding layer and the randomly initialized decoder is avoided entirely. The principal limitation is that feeding three identical channels does not exploit potential multi-channel information, but since MR input is inherently single-channel in this setting, no information is lost relative to the single-channel alternative. This approach therefore represents a more principled baseline for transfer learning from RGB-pretrained transformer encoders in grayscale medical imaging tasks, and should be investigated in future work before concluding that frozen encoder configurations are unsuitable for MR-to-CT translation.

Importance of Adversarial Training. A further finding of Experiment 3 is the strong impact of adversarial training on sCT quality. Both generator-only SwinV2-UNet configurations exhibit a substantial degradation compared to their Pix2Pix counterparts. This indicates that pixel-wise L_1 optimization alone is insufficient to enforce anatomically realistic MR-to-CT translation in this setting. The adversarial objective in Pix2Pix likely encourages sharper bone boundaries and more plausible air–tissue transitions as these seem to be distinctive features for the discriminator, which are not fully captured by L_1 loss minimization. Thus, increased architectural expressiveness alone does not compensate for the absence of adversarial supervision, underlining the importance of cGAN training (or more sophisticated loss functions) for this task.

Implications for Future Architectures. The comparative results suggest that residual learning strategies are particularly beneficial for MR-to-CT translation. Since ResNet-9 outperforms UNet, and the fully trainable SwinV2-UNet also surpasses the UNet baseline, a hybrid architecture combining residual connections with transformer-based global context modeling may represent a promising direction. For example, integrating residual connections into the bottleneck of a Swin-based encoder–decoder or constructing a Swin-ResNet hybrid backbone could potentially leverage both effective local feature refinement and long-range dependency modeling.

7.4 Limitations and Directions for Future Research

While this work follows a careful experimental approach, several important limitations should be noted, and they point to clear directions for future research.

2D Slice-wise modeling. The most significant technical limitation of this work is that all models operate on individual 2D axial slices, with no awareness of neighboring slices. As a result, reconstructed volumes can appear inconsistent when viewed from the side or front, which is also likely a major reason for the gap to the top SynthRAD2025 challenge submissions, most of which work in full 3D. Moving to 3D or hybrid 2.5D architectures would probably be the most impactful improvement for future work, at the cost of increased computational complexity.

Loss Formulations. Beyond architectural changes, the loss function is another lever worth exploring. This work relied on L_1 and adversarial objectives, but incorporating perceptual or feature-based losses, such as AFP losses, could improve structural realism and boundary sharpness, particularly for fine anatomical detail that pixel-wise objectives tend to smooth over. This aligns with observations from top SynthRAD2025 challenge submissions, which commonly combine multiple loss terms including perceptual loss.

Body Masking. CT-based body masks were used during training to improve the precision of voxel-wise supervision, but these masks are not available at inference time in an MRI-only setting. MR-based thresholding is considerably harder due to scanner- and protocol-dependent intensity variability, and developing a robust MRI-based masking strategy that generalizes across protocols remains an open problem, though TotalSegmentator's MR-based *body_trunc* task could be a promising starting point.

Dosimetric Evaluation and MR-based Contouring. The dosimetric evaluation is limited by its small cohort of three cases and the use of the left kidney as an artificial planning target, chosen for its consistent presence across abdominal cases rather than clinical relevance. Furthermore, anatomical delineations were derived from CT volumes via TotalSegmentator, since its skin-segmentation functionality is currently CT-only, whereas a genuine MRI-only workflow would require MR-based contouring. Future work should integrate MR-based segmentation methods to better approximate a clinically realistic treatment-planning pipeline. Related to this, the CT-based body masks used during training are unavailable at inference time in an MRI-only setting, and developing robust MR-based masking strategies remains an open problem.

Data Quality. On the data side, more thorough screening for misaligned or outlier image pairs – as applied by several top challenge teams – could reduce training noise and improve overall model quality.

Generalizability of Ablation Findings. Several architectural directions identified in Experiment 3 (see Section 6.3) warrant further investigation. A natural next step would be to train Pix2Pix and CycleGAN with a ResNet generator to test whether the gains observed for ResNet over UNet extend beyond the Pix2Pix-Abdomen results. On the transformer side, extending the idea behind SwinV2-UNet to a SwinV2-ResNet hybrid – where Swin Transformer blocks take the place of convolutional layers within a residual architecture – could combine the strengths of both approaches and is worth investigating as an alternative to the current UNet-like Swin formulation. Additionally, larger Swin model variants and more deliberate encoder-freezing strategies deserve further exploration; for example, also freezing the first patch embedding layer and replicating the grayscale MR input across all three input channels to better match the ImageNet pretraining assumptions.

Chapter 8

Conclusion

This project investigated sCT generation from MR images in a multi-center and multi-anatomical setting, examining three design dimensions: anatomical coverage of training data, MRI intensity normalization, and generator architecture. A recurring observation across all three is that the choice of translation framework shapes performance more strongly than any individual design decision within it. Pix2Pix and CUT proved more robust across varying conditions, while CycleGAN was consistently more sensitive – likely because the default training data augmentation of data flipping was destructive in this setting with non-symmetric, paired data.

On generalization versus specialization, joint fullbody training is feasible for Pix2Pix and CUT, but the modest and consistent advantage of region-specific models in most settings suggests that accommodating all anatomical regions within a single parameter space is not costless.

On normalization, tissue-informed methods offered marginal benefits overall, but their effect was framework-dependent: Nyúl improved Pix2Pix while degrading CUT, possibly because its histogram alignment alters spatial contrast structure in ways that interfere with CUT's patch-based contrastive objective. More broadly, the finding that simple global normalization remained competitive suggests that adversarial training is to some degree self-regularizing with respect to input intensity scale.

On architecture, the clearest result is that adversarial training proved essential – the dramatic SSIM drop in generator-only configurations confirmed that a simple L_1 regression objective alone could not enforce structural coherence at bone boundaries and air–tissue interfaces. In light of the advanced perceptual and anatomy-aware losses used in SynthRAD (e.g., TotalSegmentator-based feature constraints), it is plausible that more sophisticated supervision beyond plain L_1 might partially mitigate this gap, even without a discriminator. Among backbones, ResNet-9 outperformed the transformer-based SwinV2-UNet and the convolution-based UNet. This finding suggests that residual connections may be better suited to this regression task than purely skip-based encoder–decoder fusion.

Overall, the project demonstrated that sCT generation for MR-only radiotherapy is influenced by data composition, intensity harmonization, and architectural design choices – and that the design choices which matter most are not always the ones that receive the most attention in the literature. While our configurations did not reach the performance of the SynthRAD2025 challenge winners, the systematic ablation of methodological components provided structured insight into where performance gains and limitations arise. These findings contribute to a more transparent understanding of design trade-offs in multi-center MR-to-CT translation and provide a foundation for future improvements toward clinically robust MR-only workflows.

Statement on Use of Generative Artificial Intelligence (GenAI) Assistance

The authors utilized generative artificial intelligence tools, including ChatGPT and Claude, to assist with text and code generation. All substantive scientific contributions, experimental design, implementation, analysis, and interpretations were conducted independently by the authors.

Appendix A

Attachments

A.1 Resampling Considerations By Body Region

This appendix provides additional quantitative justification for the region-specific resampling strategy described in Section 5.1.2. The primary objective of the resampling step was threefold: (i) to obtain a unified in-plane resolution of 256×256 suitable for all downstream models, (ii) to preserve comparable anatomical scales across body regions, and (iii) to keep the body mask as large as possible within the standardized field of view while minimizing cropping. However, these objectives are partially contradicting resulting in trade-offs.

Target Resolution of 256×256. All investigated architectures (UNet-based generators, ResNet-based generators, and transformer-based variants) were trained slice-wise and benefited from a moderate, fixed spatial resolution. A resolution of 256×256 represents a practical trade-off between spatial detail and computational efficiency: it is sufficiently high to preserve relevant anatomical structures (e.g., cortical bone, air cavities, organ boundaries), while enabling stable adversarial training on available hardware. Moreover, this resolution is commonly used in image-to-image translation frameworks, ensuring architectural compatibility without further modifications.

To ensure that similar anatomical structures appear at similar pixel scales across body regions, we harmonized the effective in-plane spacing in two groups. For thorax, abdomen, and pelvis scans, this required an intermediate standardization to 512×512 followed by a downsampling by a factor of two. For brain and head/neck scans, whose body-mask bounding boxes were already substantially smaller, direct padding or cropping to 256×256 was sufficient. This results in a discontinuity in the neck area.

Distribution of Body-Mask Bounding Boxes. Figure A.1 and Figure A.2 show the empirical distribution of body-mask bounding-box sizes (width $\times$ height) per axial slice.

For brain and head/neck cases (Figure A.1), most slices clustered well below 256×256. Consequently, direct standardization to 256×256 resulted primarily in padding rather than cropping. Only 4.43% of slices required cropping of the body mask (this does not indicate how much of the body mask is actually cropped, hence, in most cases this is assumed to be as little as a few pixels), indicating that the chosen target resolution was well aligned with the anatomical extent of these regions.

For thorax, abdomen, and pelvis cases (Figure A.2), bounding boxes frequently approached 512×512. Direct resizing to 256×256 would therefore have led to substantial cropping or strong interpolation artifacts. The adopted two-stage approach (standardization to 512×512, followed

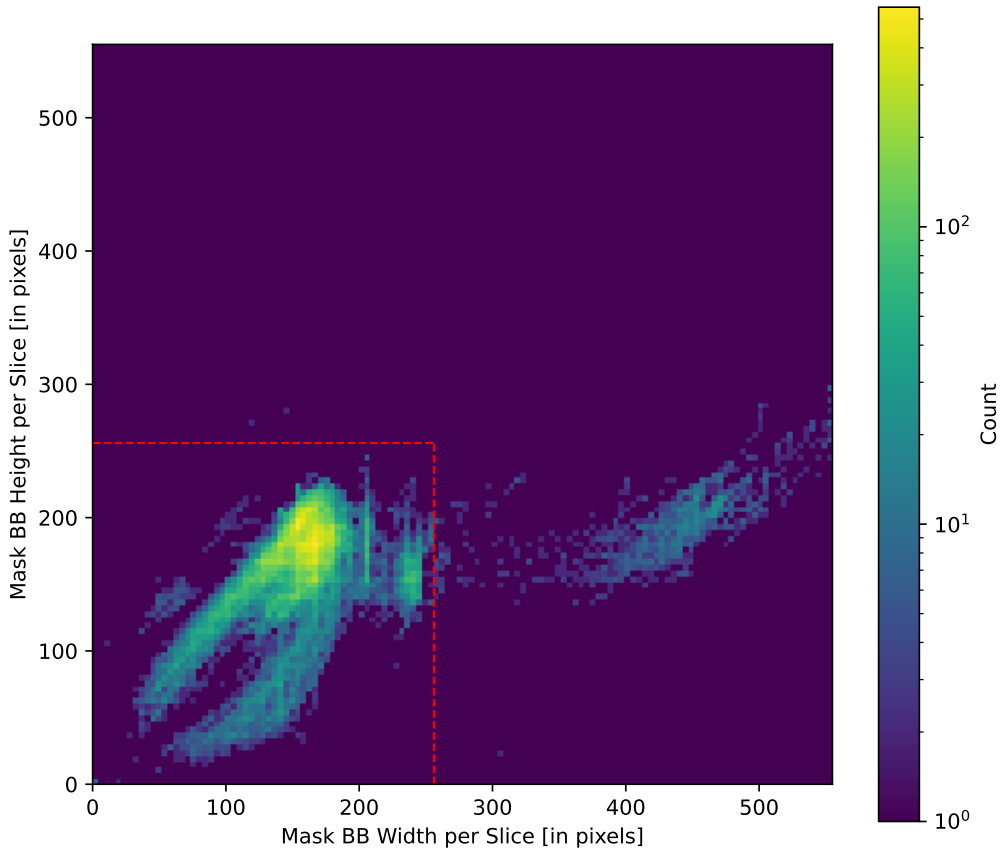


Figure A.1: HEATMAP OF BODY-MASK BOUNDING-BOX SIZES FOR HEAD/NECK AND BRAIN REGIONS. Distribution of body-mask bounding-box sizes (width $\times$ height in pixels) per axial slice for head/neck and brain regions. The majority of slices cluster well below 256×256 (dashed red line), allowing direct standardization to the target resolution without downsampling. Note: Count is in log-space.

by downsampling by a factor of two) ensured that the body mask occupied as much of the final 256×256 canvas as possible. In this configuration, only 2.24% of slices exhibited cropped masks.

Choice of Downsampling Factor and Interpolation. For the larger body regions, a downsampling factor of two was deliberately chosen. This integer scaling preserves sharp boundaries and avoids interpolation-induced blurring (when downsampling / interpolating by a non-integer factor), which is particularly important for CT bone edges and air–tissue interfaces that strongly influence HU distributions and adversarial training stability.

An alternative configuration with a downsampling factor of 1.8 would have allowed larger input structures for thorax, abdomen, and pelvis; however, this setting would result in 9.77% of slices with cropped masks and required non-integer interpolation, introducing visible smoothing effects. Given the importance of edge fidelity and structural sharpness for sCT generation, the factor-of-two strategy was preferred.

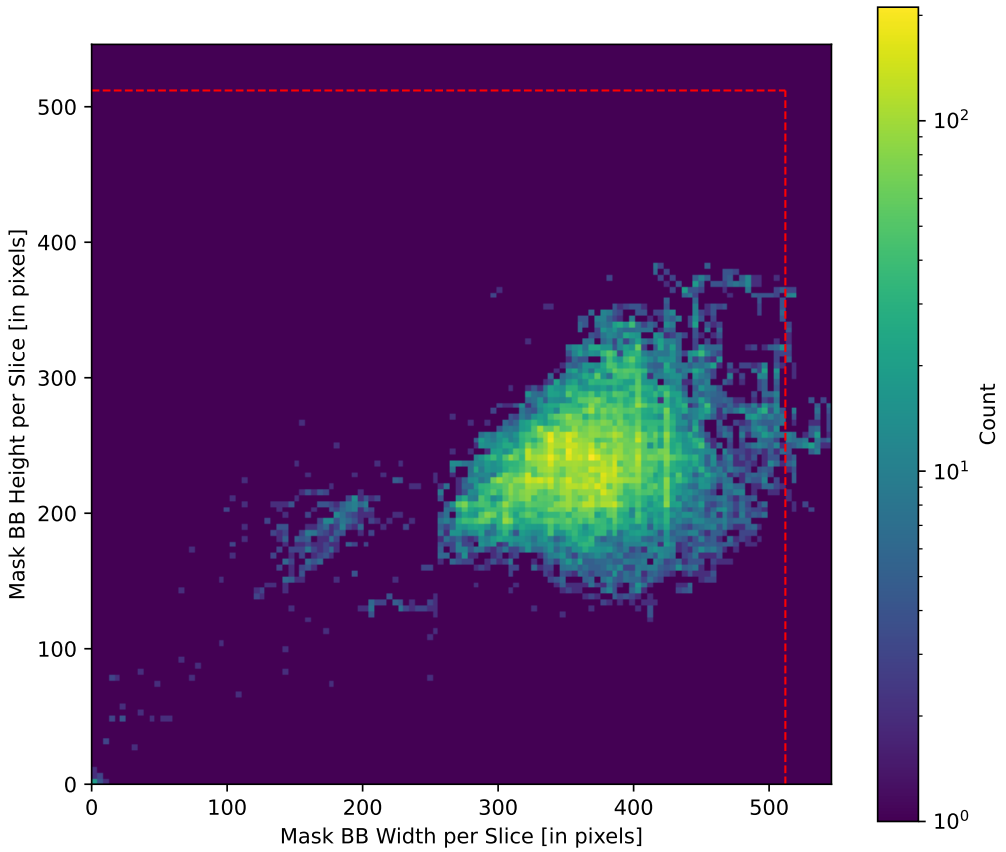


Figure A.2: Distribution of Body-Mask Bounding-Box Sizes for Thorax, Abdomen, and Pelvis Regions. Heatmap of body-mask bounding-box sizes (width $\times$ height in pixels) per axial slice for thorax, abdomen, and pelvis regions. Bounding boxes extend up to 512×512 (dashed red line), motivating the two-stage resampling approach where images are first standardized to 512×512 and then downsampled to 256×256 via nearest-neighbor interpolation. Note: Count is in log-space.

A.2 Training Configurations

This appendix summarizes the training configurations used throughout the project. Unless explicitly stated otherwise in Chapter 6 (e.g., changes in data composition, MR normalization strategy, or generator architecture), the hyperparameters and optimization settings listed below were kept fixed across all experiments to ensure comparability. The listings represent the full training options as saved by the Pix2Pix, CycleGAN and CUT training runs from the respective original implementations, provided for reproducibility purposes.

Pix2Pix Following training options have been used for all Pix2Pix (baseline with UNet-256) experiments.

```
---------------- Options ---------------
           batch_size: 1
                beta1: 0.5
```

```
       checkpoints_dir: <checkpoints> [default: ./checkpoints]
        continue_train: False
            crop_size: 256
             dataroot: <dataroot> [default: None]
          dataset_mode: aligned
             direction: AtoB
          display_freq: 400
       display_winsize: 256
                epoch: latest
           epoch_count: 1
              gan_mode: vanilla
             init_gain: 0.02
             init_type: normal
              input_nc: 1  [default: 3]
               isTrain: True  [default: None]
             lambda_L1: 100.0
             load_iter: 0  [default: 0]
             load_size: 286
                    lr: 0.0002
        lr_decay_iters: 50
            lr_policy: linear
      max_dataset_size: inf
                model: pix2pix  [default: cycle_gan]
             n_epochs: 50  [default: 100]
       n_epochs_decay: 0  [default: 100]
           n_layers_D: 3
                name: pix2pix [default: experiment_name]
                ndf: 64
               netD: basic
               netG: unet_256
                ngf: 64
            no_dropout: False
              no_flip: False
              no_html: True  [default: False]
                norm: batch
          num_threads: 4
             output_nc: 1  [default: 3]
                phase: train
            pool_size: 0
           preprocess: None  [default: resize_and_crop]
           print_freq: 100
         save_by_iter: False
       save_epoch_freq: 1  [default: 5]
      save_latest_freq: 5000
        serial_batches: False
               suffix:
      update_html_freq: 1000
            use_wandb: False
              verbose: False
    wandb_project_name: CycleGAN-and-pix2pix
----------------- End -------------------
```

CycleGAN Following training options have been used for all CycleGAN (baseline with UNet-256) experiments.

```
----------------- Options ---------------
           batch_size: 1
                beta1: 0.5
       checkpoints_dir: <checkpoints> [default: ./checkpoints]
        continue_train: False
            crop_size: 256
             dataroot: <dataroot> [default: None]
          dataset_mode: unaligned
             direction: AtoB
          display_freq: 400
       display_winsize: 256
                epoch: latest
           epoch_count: 1
              gan_mode: lsgan
             init_gain: 0.02
             init_type: normal
```

```
            input_nc: 1  [default: 3]
             isTrain: True  [default: None]
            lambda_A: 10.0
            lambda_B: 10.0
       lambda_identity: 0.5
            load_iter: 0  [default: 0]
            load_size: 286
                 lr: 0.0002
        lr_decay_iters: 50
            lr_policy: linear
      max_dataset_size: inf
               model: cycle_gan
           n_epochs: 50  [default: 100]
        n_epochs_decay: 0  [default: 100]
           n_layers_D: 3
                name: cyclegan  [default: experiment_name]
                ndf: 64
                netD: basic
                netG: unet_256  [default: resnet_9blocks]
                ngf: 64
           no_dropout: True
              no_flip: False
              no_html: False
                norm: instance
          num_threads: 4
            output_nc: 1  [default: 3]
               phase: train
            pool_size: 50
           preprocess: None  [default: resize_and_crop]
           print_freq: 100
         save_by_iter: False
       save_epoch_freq: 1  [default: 5]
      save_latest_freq: 5000
        serial_batches: False
               suffix:
      update_html_freq: 1000
            use_wandb: False
              verbose: False
     wandb_project_name: CycleGAN-and-pix2pix
----------------- End -------------------
```

CUT Following training options have been used for all CUT (baseline with ResNet-9) experiments.

```
----------------- Options ---------------
             CUT_mode: CUT
           batch_size: 1
                beta1: 0.5
                beta2: 0.999
       checkpoints_dir: <checkpoints> [default: ./checkpoints]
        continue_train: False
            crop_size: 256
             dataroot: <dataroot> [default: placeholder]
          dataset_mode: unaligned
            direction: AtoB
          display_env: main
         display_freq: 400
           display_id: None
         display_ncols: 4
          display_port: 8097
        display_server: http://localhost
       display_winsize: 256
            easy_label: experiment_name
                epoch: latest
          epoch_count: 1
       evaluation_freq: 5000
      flip_equivariance: False
             gan_mode: lsgan
              gpu_ids: 0
           init_gain: 0.02
           init_type: xavier
```

```
             input_nc: 1  [default: 3]
              isTrain: True  [default: None]
            lambda_GAN: 1.0
            lambda_NCE: 1.0
             load_size: 286
                    lr: 0.0002
          lr_decay_iters: 50
              lr_policy: linear
        max_dataset_size: inf
                 model: cut
               n_epochs: 50  [default: 200]
         n_epochs_decay: 0  [default: 200]
            n_layers_D: 3
                 name: cut  [default: experiment_name]
               nce_T: 0.07
              nce_idt: True
nce_includes_all_negatives_from_minibatch: False
            nce_layers: 0,4,8,12,16
                 ndf: 64
               netD: basic
               netF: mlp_sample
              netF_nc: 256
                netG: resnet_9blocks
                 ngf: 64
           no_antialias: False
        no_antialias_up: False
            no_dropout: True
              no_flip: False
              no_html: True  [default: False]
                normD: instance
                normG: instance
            num_patches: 256
            num_threads: 4
             output_nc: 1  [default: 3]
                phase: train
             pool_size: 0
            preprocess: None  [default: resize_and_crop]
        pretrained_name: None
            print_freq: 100
        random_scale_max: 3.0
           save_by_iter: False
        save_epoch_freq: 1  [default: 5]
        save_latest_freq: 5000
          serial_batches: False
stylegan2_G_num_downsampling: 1
                suffix:
        update_html_freq: 1000
               verbose: False
---------------- End -------------------
```

A.3 Experiment 1: By Body Region

This section provides the detailed results of Experiment 1 broken down by anatomical region. While the main text reports metrics aggregated over all regions, Figure A.3 visualizes the per-region distributions for MAE, MSE, PSNR, and SSIM based on our internal image similarity evaluation.

Overall trends across regions. Although absolute performance levels varied considerably between anatomical regions, the global trends described in Section 6.1 remained largely consistent. In most regions and across most metrics, region-specific training performed slightly better than the corresponding fullbody model. Furthermore, Pix2Pix consistently outperformed CUT in the majority of settings, while CycleGAN showed clearly inferior performance across nearly all regions and metrics. These tendencies were reflected both in error-based metrics (MAE, MSE) and

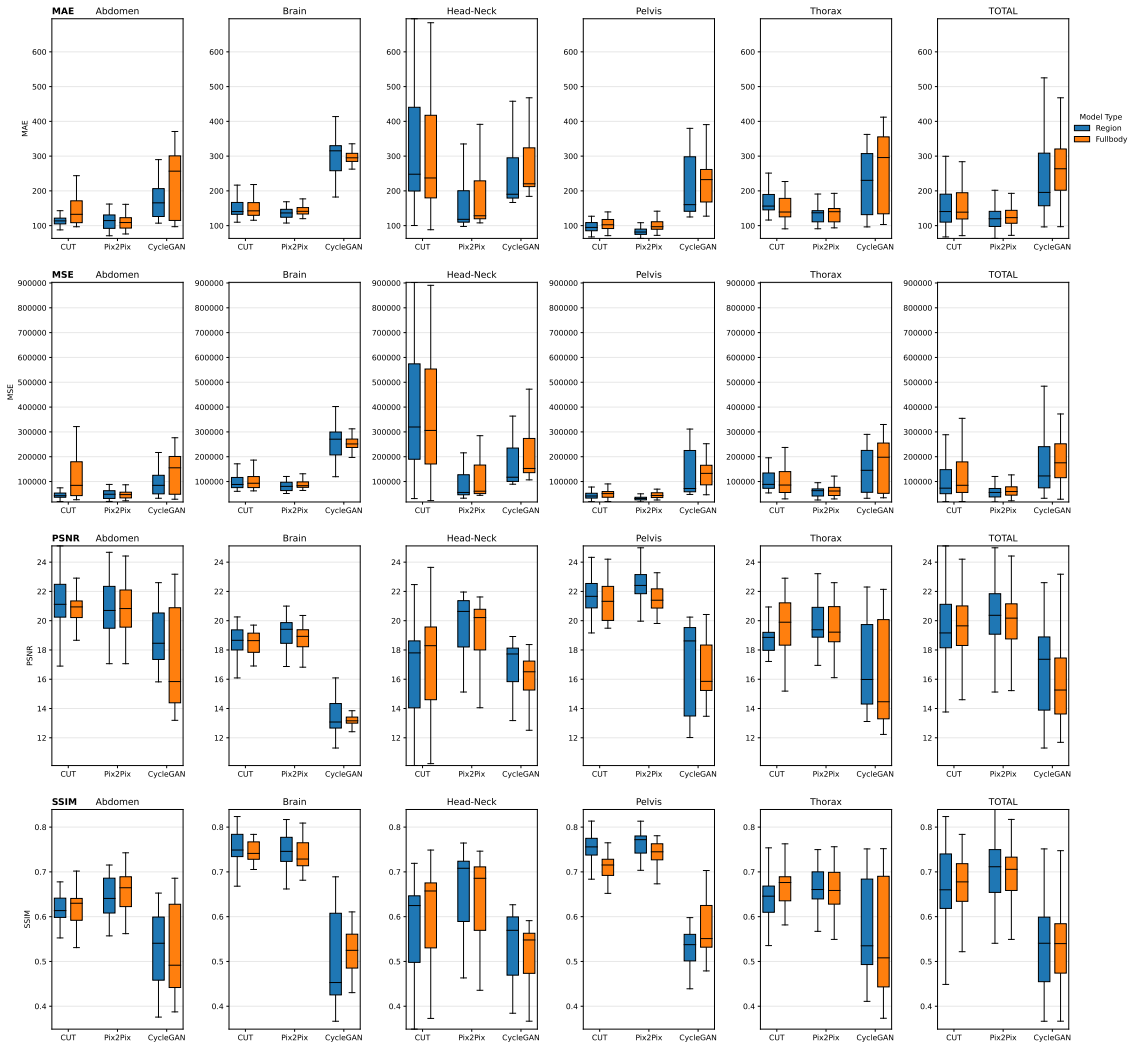


Figure A.3: Experiment 1 Results by Anatomical Region Based on Own Image Similarity Evaluation. Boxplot of Experiment 1 results by anatomical region as well as aggregated ("TOTAL"). For each framework (CUT, Pix2Pix, CycleGAN), one jointly trained fullbody model is shown alongside five separately trained region-specific models (abdomen, brain, head-neck, pelvis, thorax). Lower is better for MAE and MSE; higher is better for PSNR and SSIM. Evaluation on test set (141 volumes across all anatomical regions).

perceptual metrics (PSNR, SSIM), indicating that the aggregated conclusions were not driven by a single anatomical site.

Region-specific observations. The pelvis region exhibited comparatively favorable performance for Pix2Pix and CUT, characterized by lower MAE/MSE and higher PSNR/SSIM relative to more heterogeneous regions. In contrast, the head-and-neck region showed substantially larger variability and overall reduced performance, particularly visible in the wide interquartile ranges and outliers for MAE and MSE. This likely reflects the anatomical complexity of the region, including fine bone–air interfaces and small high-contrast structures.

For the brain region, performance was relatively stable with moderate dispersion, and differences between fullbody and region-specific models were small. The thorax demonstrated increased variability, likely related to lung tissue and sharp air–tissue transitions, which seem challenging for 2D slice-wise models.

Consistency of fullbody vs. region-specific training. Across most anatomical regions, region-specific models achieved equal or slightly improved median performance compared to the fullbody models. However, the magnitude of these differences was generally modest for Pix2Pix and CUT. CycleGAN showed a higher sensitivity to joint training, with the fullbody configuration often exhibiting higher errors and lower perceptual scores than the region-specific variants.

Taken together, the per-region analysis confirms the conclusions drawn from the aggregated results: (i) region-specific training provides a small but consistent advantage in most cases, (ii) Pix2Pix represents the strongest framework among the evaluated methods, and (iii) CycleGAN underperforms substantially across anatomical regions. While anatomical difficulty modulates absolute performance, the relative ranking between frameworks and training regimes remains largely stable.

A.4 Experiment 1: Effect of Mask Quality for Pix2Pix

This section analyzes the substantial performance drop observed for Pix2Pix between the internal image similarity evaluation (MAE $= 137.19$ for the fullbody model) and the external evaluation in SynthRAD2025 (MAE $= 184.91$ for the fullbody model). The discrepancy was particularly pronounced for MAE and therefore warranted a dedicated analysis of potential contributing factors.

Table A.1 reports the performance of the same fullbody Pix2Pix model on the three SynthRAD2025 anatomical regions (abdomen, head-neck, thorax), evaluated once using the CT-derived body masks employed in our experiments and once using the original, comparatively broad masks provided by the challenge organizers (respectively applied to input and output). Approximately 20% of the observed difference could be attributed to the restriction to the anatomical subset, and roughly 40% to differences in mask definition. The remaining discrepancy (approximately 40%) was likely caused by additional factors, including dataset-specific variations.

Qualitative inspection revealed that the broader challenge masks frequently included anatomical structures such as extremities (e.g., arms) that were not consistently present in both modalities (MR and CT). Our CT-derived masks were designed to minimize such mismatches by restricting evaluation to regions with valid cross-modality correspondence. In contrast, the original SynthRAD masks did not account for this issue, which led to inflated voxel-wise errors and, consequently, markedly degraded quantitative metrics.

Table A.1: EXPERIMENT 1 RESULTS BY MASK TYPE. Experiment 1 results aggregated over the three SynthRAD2025 anatomical regions (abdomen, head-neck, thorax) to assess effect of mask quality on metrics based on own image similarity evaluation (mean ± std), test set (87 volumes).

Model	Body Masks	MAE	MSE	PSNR	SSIM
Pix2Pix (Fullbody)	Baseline (CT-derived)	146.99 ± 71.40	83'553 ± 81'055	19.88 ± 1.99	0.655 ± 0.065
Pix2Pix (Fullbody)	SynthRAD original masks	163.34 ± 45.61	92'284 ± 42'865	17.94 ± 1.94	0.652 ± 0.066

A.5 Experiment 2: Extended Results

This section provides results for Experiment 2 with more granularity. Table A.2 reports the full quantitative results of the different npeaks configurations of Experiment 2 on the abdominal test set, including mean and standard deviation for all four image similarity metrics (MAE, MSE, PSNR, SSIM) across all frameworks. This allows direct comparison between N-Peaks configurations under otherwise unchanged training conditions.

Figure A.4 presents the Experiment 2 image similarity results on abdominal data as boxplots, broken down by framework and N-Peaks configuration. The figure complements Table A.2 by visualizing the spread and distribution of per-patient metrics, allowing identification of outliers and assessment of result stability across configurations.

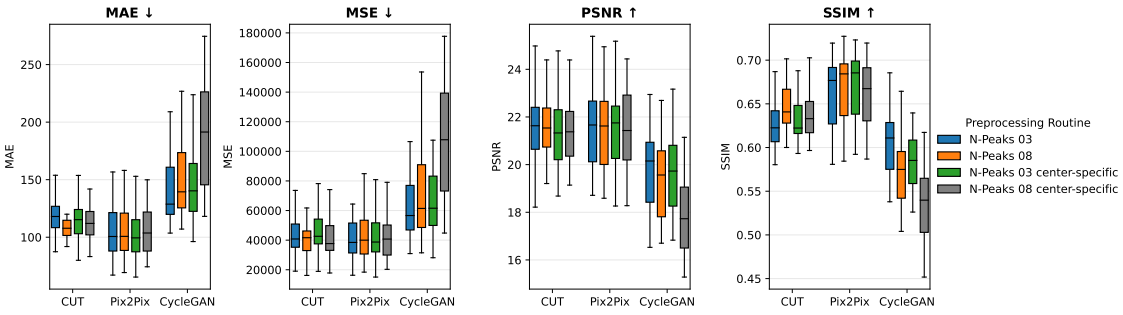


Figure A.4: EXPERIMENT 2 RESULTS ON ABDOMINAL DATA BASED ON OWN IMAGE SIMILARITY EVALUATION (FULL), TEST SET (27 VOLUMES).

Table A.3 presents the results of the model-based approach *Separate 1st Layer* broken down by treatment center of the data. It indicates that some models were not able to learn meaningful feature extractions for center C.

Table A.2: EXPERIMENT 2 RESULTS ON ABDOMINAL DATA BASED ON OWN IMAGE SIMILARITY EVALUATION. Experiment 2 results on abdominal data based on own image similarity evaluation (mean ± std), test set (27 volumes).

Framework	Preprocessing Routine	MAE	MSE	PSNR	SSIM
CUT	N-Peaks 03	121.40 ± 25.60	46'555 ± 24'330	**21.75 ± 1.86**	0.626 ± 0.027
CUT	N-Peaks 08	113.56 ± 25.27	**44'672 ± 23'800**	21.71 ± 1.65	0.646 ± 0.028
CUT	N-Peaks 03 center-specific	120.69 ± 35.32	51'488 ± 31'836	21.30 ± 2.05	0.629 ± 0.030
CUT	N-Peaks 08 center-specific	119.57 ± 36.74	47'945 ± 33'792	21.16 ± 1.69	0.633 ± 0.033
CycleGAN	N-Peaks 03	141.39 ± 31.43	65'766 ± 29'067	19.85 ± 1.70	0.603 ± 0.036
CycleGAN	N-Peaks 08	149.97 ± 30.28	70'867 ± 28'632	19.44 ± 1.68	0.571 ± 0.037
CycleGAN	N-Peaks 03 center-specific	147.00 ± 34.52	70'421 ± 32'339	19.64 ± 1.64	0.584 ± 0.034
CycleGAN	N-Peaks 08 center-specific	191.99 ± 46.03	107'480 ± 39'909	17.79 ± 1.68	0.533 ± 0.045
Pix2Pix	N-Peaks 03	106.22 ± 27.49	44'813 ± 23'263	21.59 ± 1.71	0.662 ± 0.040
Pix2Pix	N-Peaks 08	106.85 ± 28.23	45'462 ± 23'803	21.51 ± 1.73	0.666 ± 0.043
Pix2Pix	N-Peaks 03 center-specific	**105.28 ± 29.34**	44'904 ± 25'176	21.65 ± 1.84	**0.671 ± 0.039**
Pix2Pix	N-Peaks 08 center-specific	108.73 ± 29.19	46'050 ± 25'117	21.48 ± 1.70	0.661 ± 0.039

Table A.3: SEPARATE FIRST LAYER RESULTS BY TREATMENT CENTER ON ABDOMINAL DATA. Results based on own image similarity evaluation (mean ± std), test set (number of training samples for the respective center: $n_A = 10, \quad n_B = 14, \quad n_C = 3$).

model	center	MAE	MSE	PSNR	SSIM
CUT	A	104.50 ± 17.15	34'609 ± 12'051	22.47 ± 1.62	0.637 ± 0.027
CUT	B	127.67 ± 14.85	56'673 ± 13'081	20.19 ± 1.09	0.606 ± 0.026
CUT	C	869.37 ± 22.40	847'066 ± 33'266	8.47 ± 0.25	0.251 ± 0.030
Pix2Pix	A	92.23 ± 19.9	33'461 ± 12'013	22.17 ± 1.39	0.676 ± 0.036
Pix2Pix	B	132.16 ± 24.86	64'474 ± 20'267	19.64 ± 1.24	0.606 ± 0.045
Pix2Pix	C	130.77 ± 71.60	70'102 ± 65'385	21.35 ± 2.46	0.674 ± 0.067
CycleGAN	A	122.05 ± 23.98	48'898 ± 16'822	20.82 ± 1.52	0.600 ± 0.033
CycleGAN	B	261.23 ± 31.10	175'194 ± 29'605	14.99 ± 0.85	0.436 ± 0.039
CycleGAN	C	370.02 ± 414.6	311'525 ± 438'333	16.26 ± 6.98	0.499 ± 0.193

A.6 Experiment 2: N-Peaks configurations

This section presents the visual inspection results for all twelve N-Peaks configurations, which were assessed based on their tissue intensity distributions prior to selecting a subset of four configurations for downstream model training evaluation.

Figure A.5 overlays the intensity kernel density estimates of all twelve N-Peaks configurations on the abdominal set after normalization. The close alignment of the curves across configurations indicates that all variants successfully map intensities into the $[0, 1]$ range with broadly similar distributional shape. Subtle differences are most visible at the low-intensity peak near zero and at the upper tail. The overall similarity of the curves suggests that the macro-level intensity distribution is relatively robust to the specific parameter choices, with differences in downstream sCT quality likely driven by finer tissue-level alignment rather than global distributional shifts.

Figure A.6 focuses on liver tissue intensity distributions. Since liver serves as one of the primary anchoring peaks in the N-Peaks method, all configurations produce similarly well-defined peaks across the three centers, with no substantial differences between configurations. This consistency holds despite the smaller Center C cohort and the differing field strengths across acquisition sites.

Figure A.7 focuses on fat tissue intensity distributions. In contrast to the liver, the fat mask does not produce clear peaks under any configuration, regardless of whether global or center-specific scope is used. Centers A and C exhibit a slight peak, but Center B shows no discernible peak at all. The failure to establish a reliable fat anchor for Center B likely limits the overall

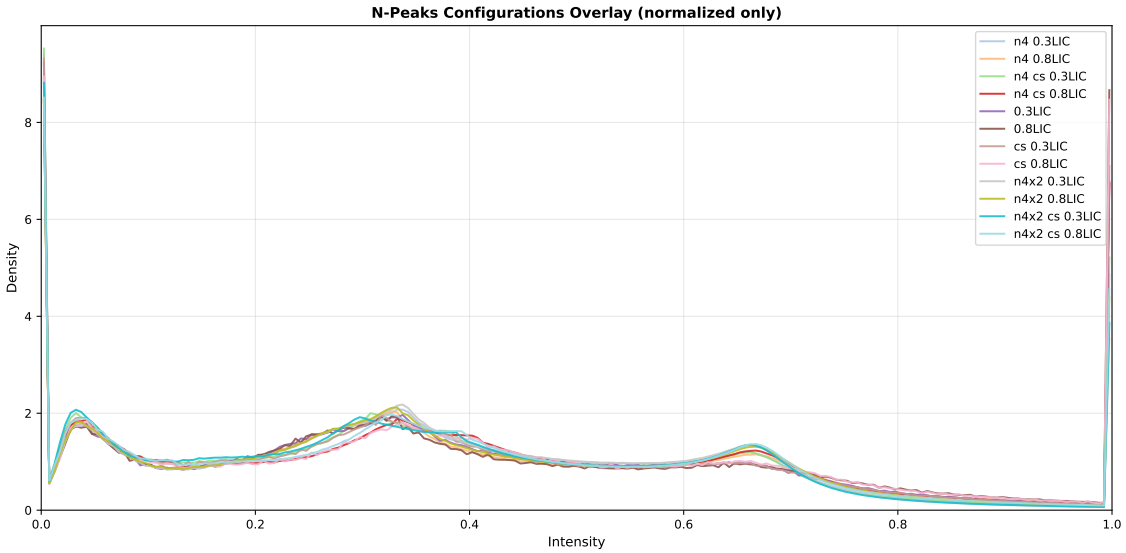


Figure A.5: OVERLAY OF INTENSITY DISTRIBUTIONS ACROSS N-PEAKS CONFIGURATIONS. Overlay of intensity kernel density estimates across all twelve N-Peaks configurations on the abdominal training set, showing the effect of LIC threshold ($q \in \{0.3, 0.8\}$), scope (global vs. center-specific), and N4 bias field correction iterations (none, $\times 1$, $\times 2$).

effectiveness of the N-Peaks normalization in this dataset.

Figures A.8–A.11 show the N-Peaks visualizer output for patient ABB042, illustrating the mask extraction, LIC computation, and histogram peak identification steps of the N-Peaks algorithm. Each figure corresponds to a specific tissue mask (liver or fat) and LIC threshold ($q \in \{0.3, 0.8\}$). The visualizations include the MR slice with the tissue mask overlay, the LIC map, and the resulting intensity histogram with the identified peak used for normalization. The LIC map classifies each voxel within the mask as either homogeneous or inhomogeneous based on the local intensity coherence criterion: homogeneous voxels exhibit intensities consistent with their local neighborhood and are retained for peak estimation, while inhomogeneous voxels are excluded to avoid corrupting the peak location. For the liver mask, the retained homogeneous voxels produce a sharp, well-defined histogram peak, reflecting the relatively uniform signal intensity of liver tissue. For the fat mask, the peak is considerably broader and less pronounced, consistent with the heterogeneous nature of the fat tissue discussed above. This difference in tissue homogeneity explains why the liver produces a much sharper normalization anchor than the fat mask. These outputs were used to qualitatively assess the stability and plausibility of the peak detection across patients and configurations.

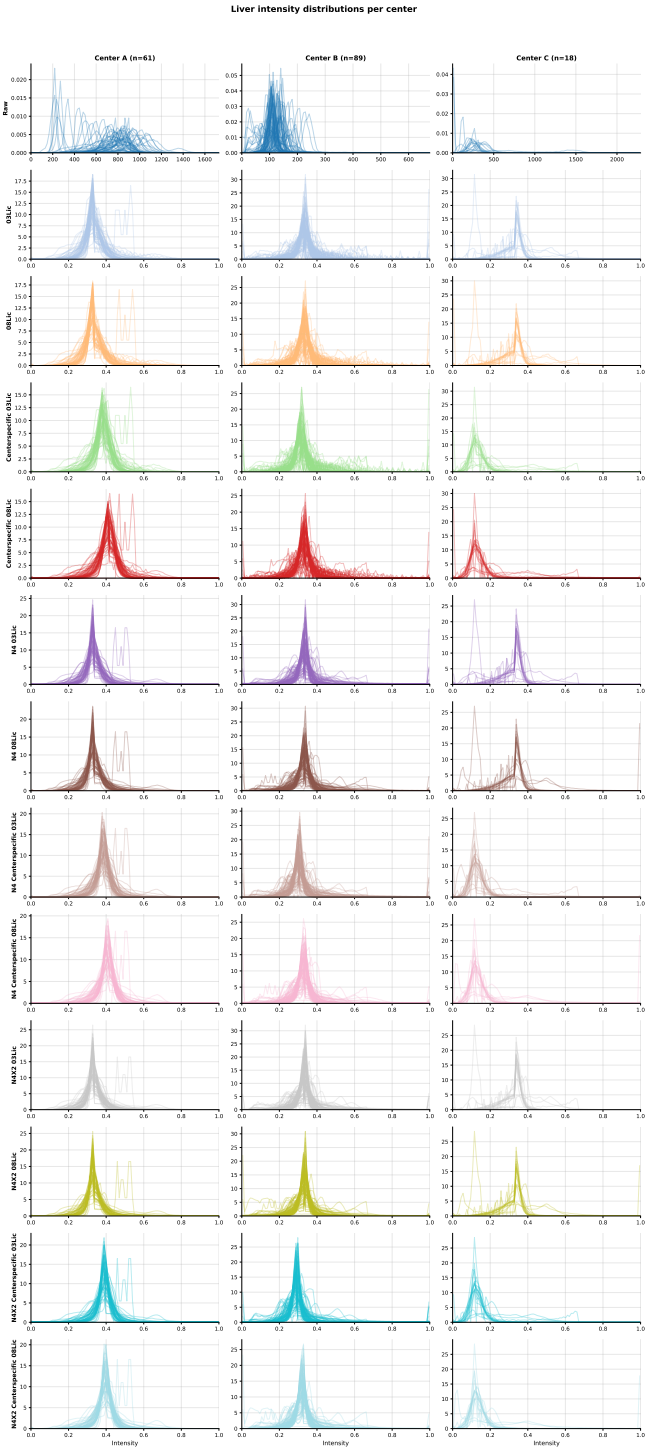


Figure A.6: PER-CENTER LIVER INTENSITY DISTRIBUTIONS ACROSS N-PEAKS CONFIGURATIONS. Per-center liver intensity distributions across all twelve N-Peaks configurations. Columns correspond to Center A (61 volumes), Center B (89 volumes), and Center C (18 volumes). The top row shows raw intensities; subsequent rows show each N-Peaks configuration varying LIC threshold ($q \in \{0.3, 0.8\}$), scope (global vs. center-specific), and N4 iterations (none, $\times 1$, $\times 2$).

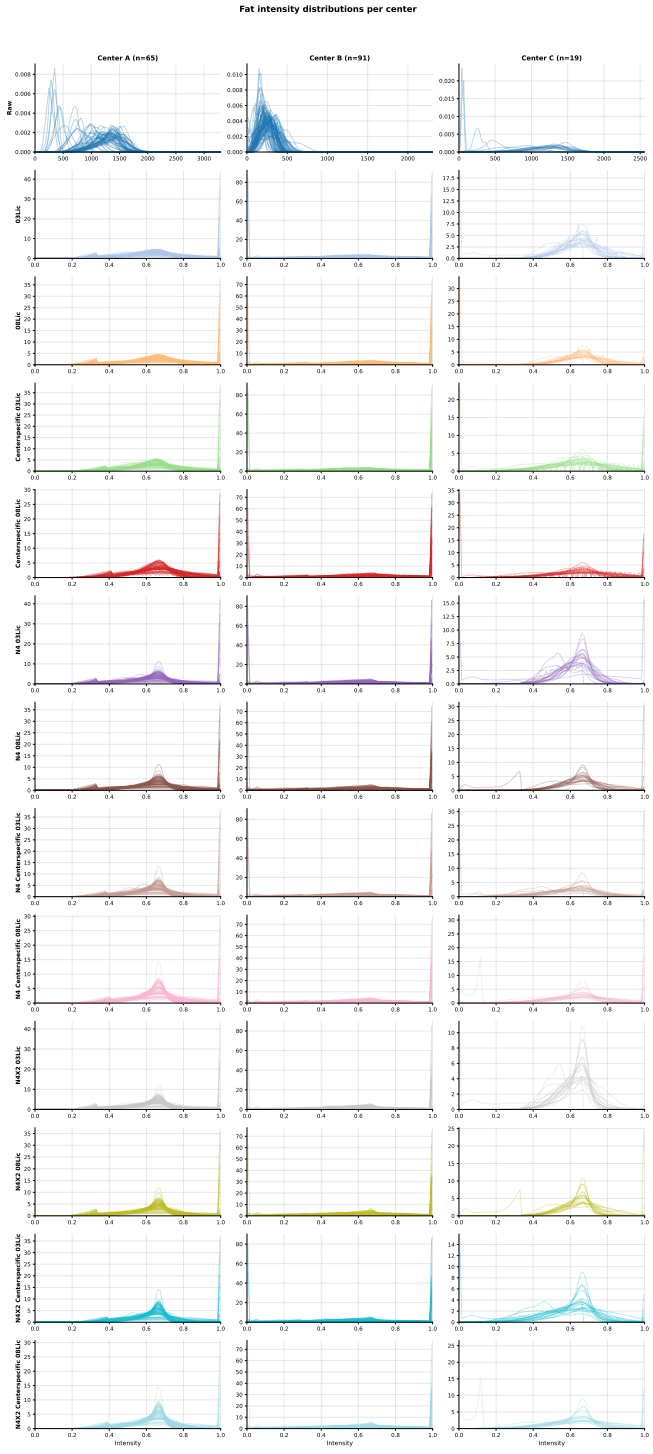


Figure A.7: PER-CENTER FAT INTENSITY DISTRIBUTIONS ACROSS N-PEAKS CONFIGURATIONS. Per-center fat intensity distributions across all twelve N-Peaks configurations. Columns correspond to Center A (65 volumes), Center B (91 volumes), and Center C (19 volumes). The top row shows raw intensities; subsequent rows show each N-Peaks configuration varying LIC threshold ($q \in \{0.3, 0.8\}$), scope (global vs. center-specific), and N4 iterations (none, $\times 1$, $\times 2$).

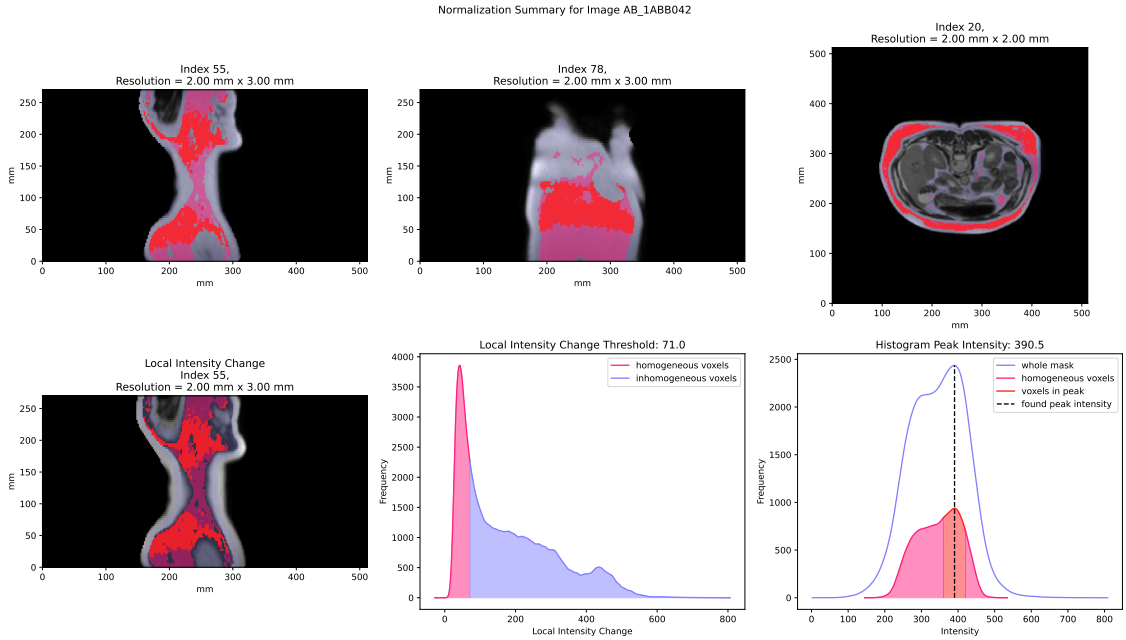


Figure A.8: N-PEAKS VISUALIZER OUTPUT FOR LIVER MASK WITH LIC THRESHOLD 0.3. N-Peaks visualizer output for the liver mask with LIC threshold $q = 0.3$.

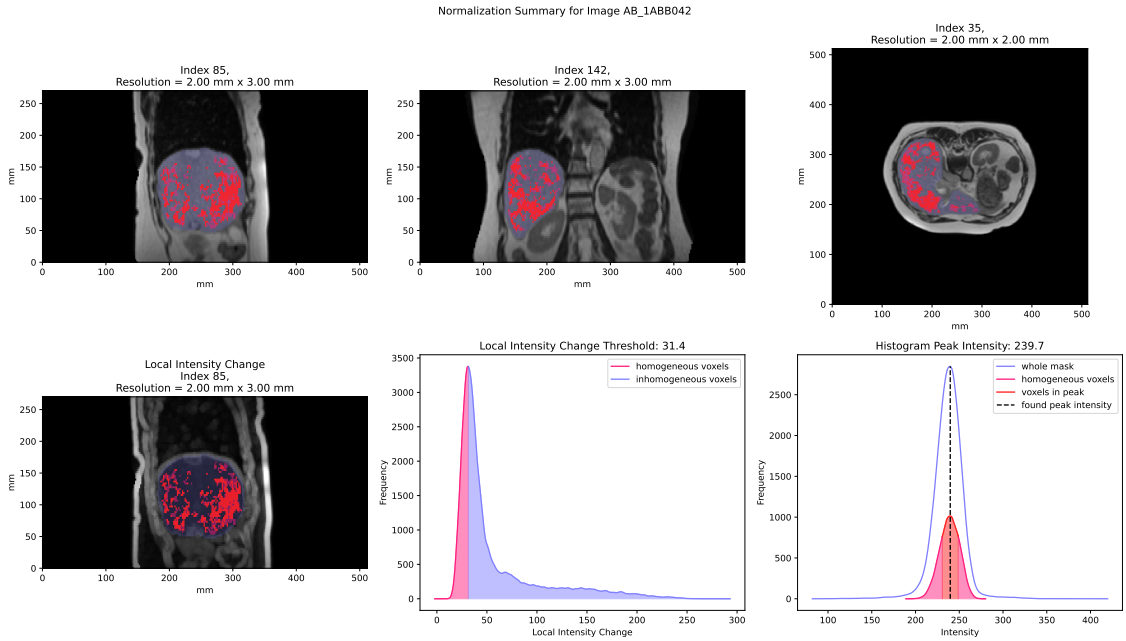


Figure A.9: N-PEAKS VISUALIZER OUTPUT FOR FAT MASK WITH LIC THRESHOLD 0.3. N-Peaks visualizer output for the fat mask with LIC threshold $q = 0.3$.

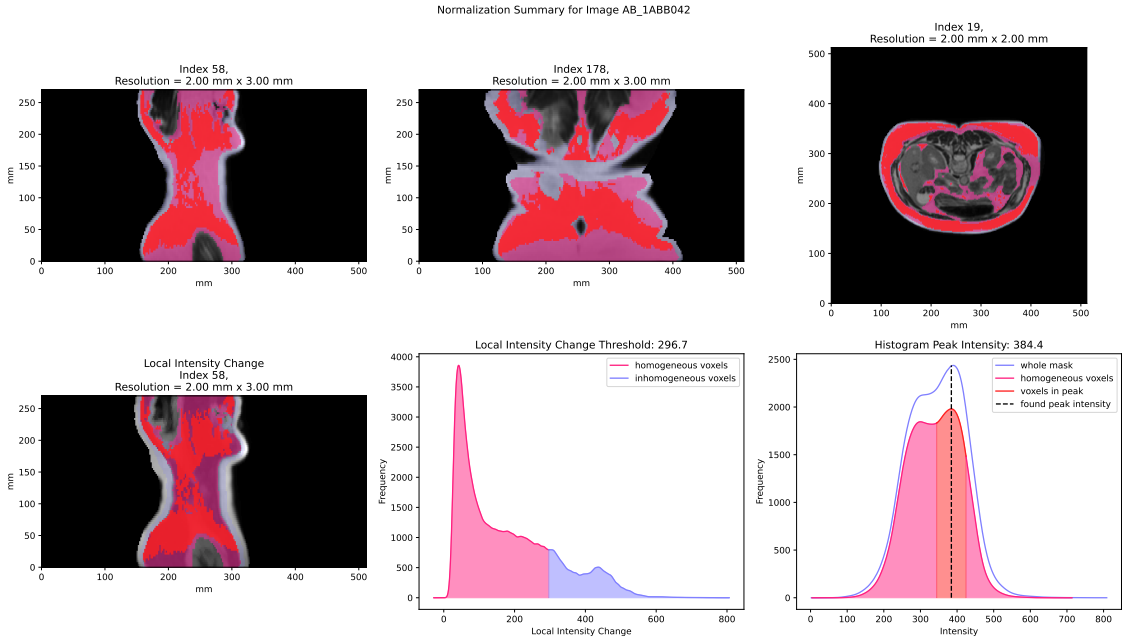


Figure A.10: N-PEAKS VISUALIZER OUTPUT FOR LIVER MASK WITH LIC THRESHOLD 0.8. N-Peaks visualizer output for the liver mask with LIC threshold $q = 0.8$.

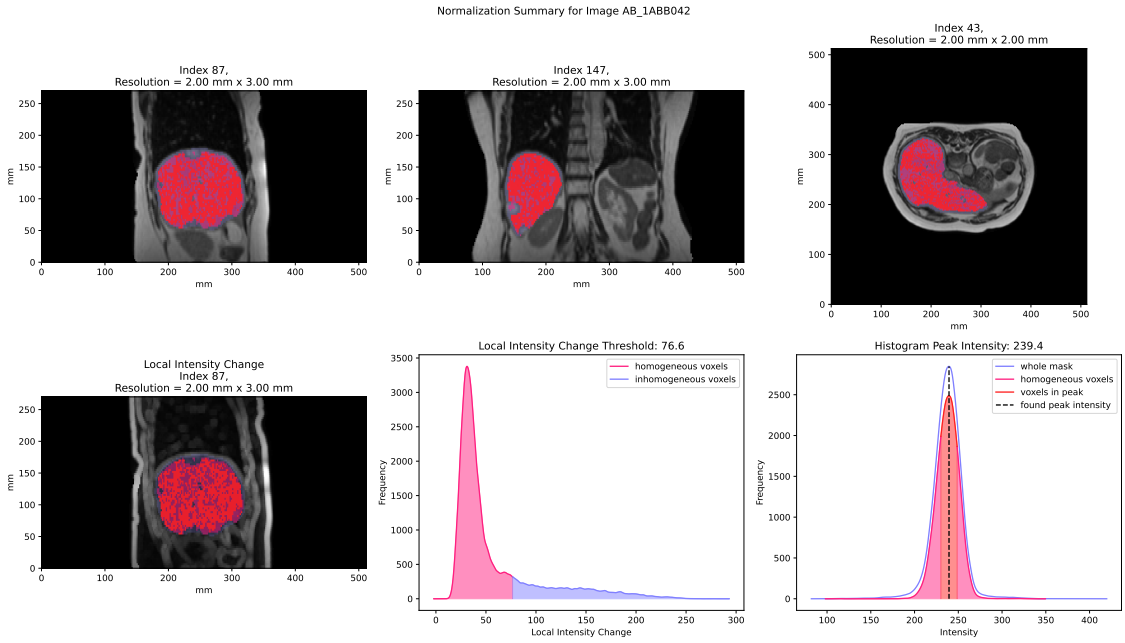


Figure A.11: N-PEAKS VISUALIZER OUTPUT FOR FAT MASK WITH LIC THRESHOLD 0.8. N-Peaks visualizer output for the fat mask with LIC threshold $q = 0.8$.

A.7 Experiment 2: Dosimetric Evaluation

Figures A.12, A.13, and A.14 show the DVH indicator differences between dCT- and sCT-based dose plans across all evaluated input normalization strategies for the Pix2Pix, CUT, and CycleGAN models, respectively. Due to the small sample size, no statistically significant differences between normalization strategies could be detected: Friedman test p-values exceeded the significance threshold of $p = 0.05$ for all models and all normalization configurations. Notably, for Pix2Pix and CUT, the majority of DVH indicator differences remained within the $\pm 2\%$ clinical acceptability threshold, suggesting that the choice of normalization strategy has limited dosimetric impact for these models.

Figure A.15 illustrates an exemplary dose distribution calculated on a sCT with automatically derived organ-at-risk contours obtained via TotalSegmentator.

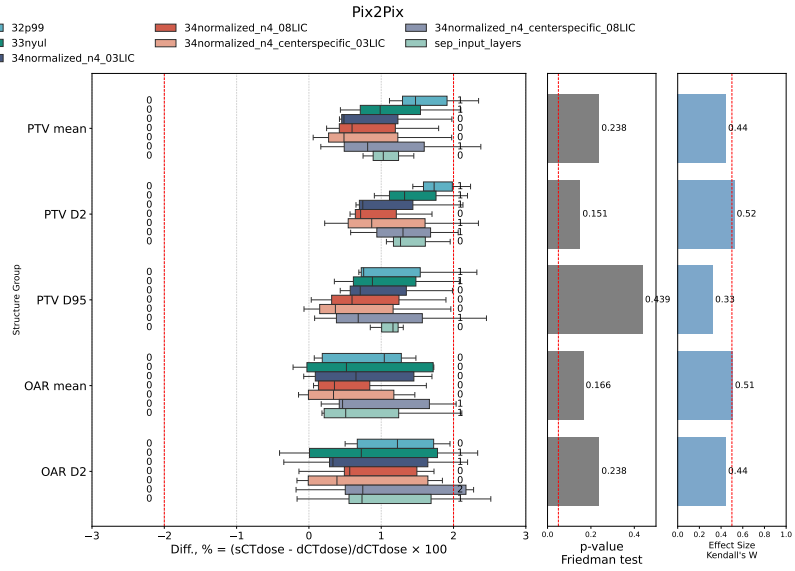


Figure A.12: NORMALIZATION STRATEGIES: PIX2PIX. DVH indicator differences between dCT and sCT-based dose plans are shown for various input normalization strategies applied to the Pix2Pix model. Outliers exceeding the $\pm 2\%$ threshold are counted next to the vertical red dashed lines. The middle panels show Friedman test p-values, with the significance threshold at $p = 0.05$ indicated by a dashed line. The right panels display effect sizes as Kendall's W, where values above 0.5 indicate a large effect.

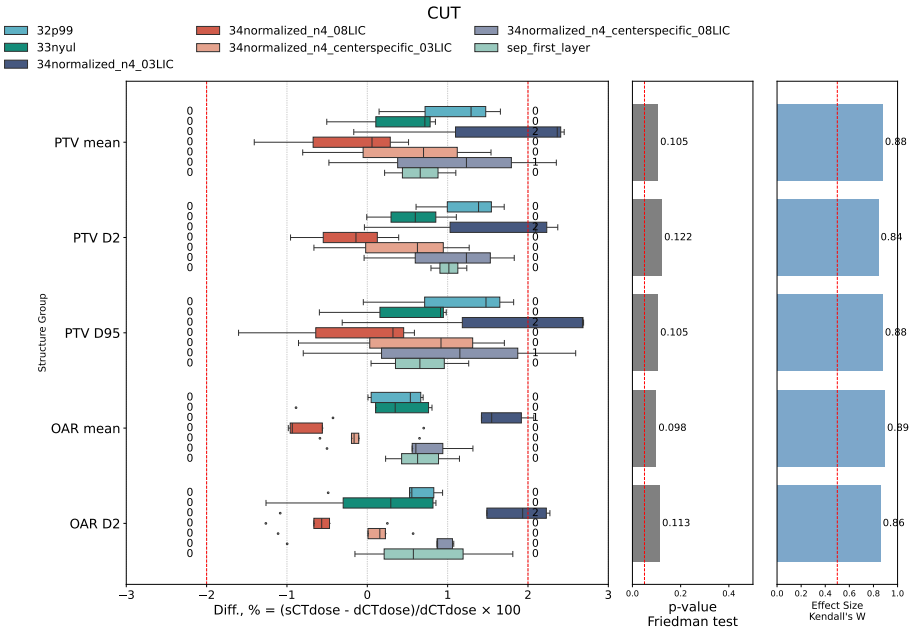


Figure A.13: NORMALIZATION STRATEGIES: CUT. DVH indicator differences between dCT and sCT-based dose plans are shown for various input normalization strategies applied to the CUT model. Outliers exceeding the $\pm 2\%$ threshold are counted next to the vertical red dashed lines. The middle panels show Friedman test p-values, with the significance threshold at $p = 0.05$ indicated by a dashed line. The right panels display effect sizes as Kendall's W, where values above 0.5 indicate a large effect.

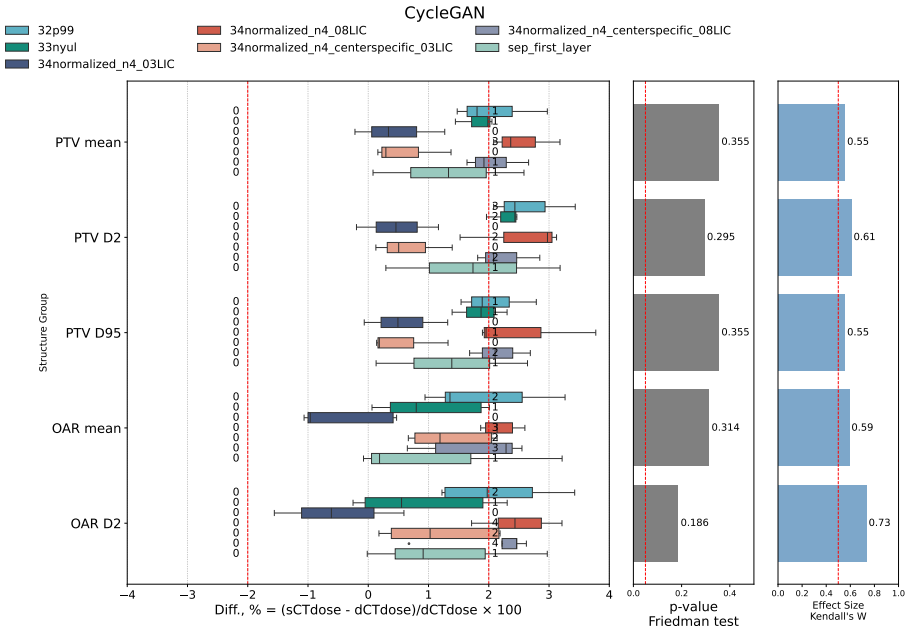


Figure A.14: NORMALIZATION STRATEGIES: CYCLEGAN. DVH indicator differences between dCT and sCT-based dose plans are shown for various input normalization strategies applied to the CycleGAN model. Outliers exceeding the $\pm 2\%$ threshold are counted next to the vertical red dashed lines. The middle panels show Friedman test p-values, with the significance threshold at $p = 0.05$ indicated by a dashed line. The right panels display effect sizes as Kendall's W, where values above 0.5 indicate a large effect.

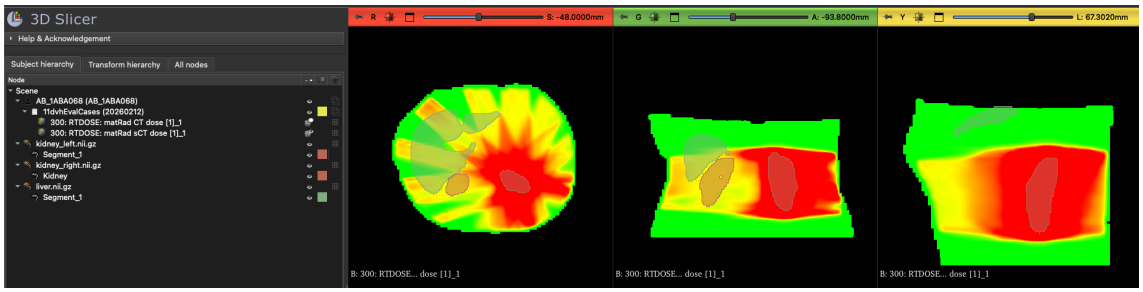


Figure A.15: DOSE DISTRIBUTION ON SCT WITH TOTALSEGMENTATOR-DERIVED STRUCTURES. Visualization in 3D Slicer of a dose distribution calculated on a sCT, shown in axial, coronal, and sagittal views. The dose is displayed as a color wash ranging from green (low dose) to red (high dose). Organ-at-risk contours for the left kidney, right kidney, and liver were automatically derived using TotalSegmentator and are overlaid as dashed outlines.

List of Figures

List of Tables

Bibliography

Akinci D'Antonoli, T., Berger, L. K., Indrakanti, A. K., Vishwanathan, N., Weiss, J., Jung, M., Berkarda, Z., Rau, A., Reisert, M., Küstner, T., Walter, A., Merkle, E. M., Boll, D. T., Breit, H.-C., Nicoli, A. P., Segeroth, M., Cyriac, J., Yang, S., and Wasserthal, J. (2025). Totalsegmentator MRI: Robust sequence-independent segmentation of multiple anatomic structures in MRI. *Radiology*, 314(2).

AlHadidi, A., LaPrade, J. C., and Kan Yeung, A. W. (2026). Basic Principles of CT and MR Imaging. *Oral and Maxillofacial Surgery Clinics of North America*, 38(1):1–13.

Boulanger, M., Nunes, J.-C., Chourak, H., Largent, A., Tahri, S., Acosta, O., De Crevoisier, R., Lafond, C., and Barateau, A. (2021). Deep learning methods to generate synthetic CT from MRI in radiotherapy: A literature review. *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB)*, 89.

Boussot, V., Hémon, C., Nunes, J.-C., and Dillenseger, J.-L. (2025). Why registration quality matters: Enhancing sCT synthesis with IMPACT-based registration. *arXiv preprint arXiv:2510.21358*.

Brou Boni, K. N. D., Klein, J., Vanquin, L., Wagner, A., Lacornerie, T., Pasquier, D., and Reynaert, N. (2020). MR to CT synthesis with multicenter data in the pelvic area using a conditional generative adversarial network. *Physics in Medicine & Biology*, 65(7):075002.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2023). Swin-Unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham. Springer Nature Switzerland.

Chen, W. Z., Xiao, Y., and Li, J. (2014). Impact of dose calculation algorithm on radiation therapy. *World Journal of Radiology*, 6(11):874–880.

Dalmaz, O., Yurt, M., and Çukur, T. (2022). ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614.

Dayarathna, S., Islam, K. T., Uribe, S., Yang, G., Hayat, M., and Chen, Z. (2024). Deep learning based synthesis of MRI, CT and PET: Review and analysis. *Medical image analysis*, 92.

Dhariwal, P. and Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Drzymala, R. E., Mohan, R., Brewster, L., Chu, J., Goitein, M., Harms, W., and Urie, M. (1991). Dose-volume histograms. *International Journal of Radiation Oncology, Biology, Physics*, 21(1):71–78.

Fetty, L., Löfstedt, T., Heilemann, G., Furtado, H., Nesvacil, N., Nyholm, T., Georg, D., and Kuess, P. (2020). Investigating conditional GAN performance with different generator architectures, an ensemble model, and different MR scanners for MR-sCT conversion. *Physics in Medicine & Biology*, 65(10):105004.

Fu, J., Yang, Y., Singhrao, K., Ruan, D., Chu, F.-I., Low, D. A., and Lewis, J. H. (2019). Deep learning approaches using 2D and 3D convolutional neural networks for generating male pelvic synthetic computed tomography from magnetic resonance imaging. *Medical Physics*, 46(9):3788–3798.

González, J. S., Longuefosse, A., Díaz Benito, M., García Martín, Á., and Baldacci, F. (2025). Deep learning-based cross-anatomy CT synthesis using adapted nnResU-Net with anatomical feature prioritized loss. *arXiv preprint arXiv:2509.22394*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2020). Generative adversarial networks. *Commun. ACM*, 63(11):139–144.

Han, B. and Tan, Z. (2025). A MR-to-CT synthesis method for SynthRAD2025 challenge. Algorithm description, SynthRAD2025 Challenge, Task 1.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hounsfield, G. N. (1980). Computed medical imaging. *Science*, 210(4465):22–28.

Hsu, S.-H., Han, Z., Hu, Y.-H., Ferguson, D., van Dams, R., Mak, R. H., Leeman, J. E., and Sudhyadhom, A. (2025). Feasibility study of a general model for synthetic CT generation in MRI-guided extracranial radiotherapy. *Biomedical Physics & Engineering Express*, 11(3):035028.

Huijben, E. M., Terpstra, M. L., Galapon, A. J., Pai, S., Thummerer, A., Koopmans, P., Afonso, M., van Eijnatten, M., Gurney-Champion, O., Chen, Z., Zhang, Y., Zheng, K., Li, C., Pang, H., Ye, C., Wang, R., Song, T., Fan, F., Qiu, J., Huang, Y., Ha, J., Sung Park, J., Alain-Beaudoin, A., Bériault, S., Yu, P., Guo, H., Huang, Z., Li, G., Zhang, X., Fan, Y., Liu, H., Xin, B., Nicolson, A., Zhong, L., Deng, Z., Müller-Franzes, G., Khader, F., Li, X., Zhang, Y., Hémon, C., Boussot, V., Zhang, Z., Wang, L., Bai, L., Wang, S., Mus, D., Kooiman, B., Sargeant, C. A., Henderson, E. G., Kondo, S., Kasai, S., Karimzadeh, R., Ibragimov, B., Helfer, T., Dafflon, J., Chen, Z., Wang, E., Perko, Z., and Maspero, M. (2024). Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report. *Medical Image Analysis*, 97:103276.

International Commission on Radiation Units and Measurements (2010). Prescribing, recording, and reporting photon-beam intensity-modulated radiation therapy (IMRT). Technical Report 83, International Commission on Radiation Units and Measurements, Bethesda, MD.

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

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976.

Jang, S.-I., Lois, C., Thibault, E. G., Becker, J. A., Dong, Y., Normandin, M. D., Price, J. C., Johnson, K. A., Fakhri, G. E., and Gong, K. (2023). Taupetgen: Text-conditional tau pet image synthesis based on latent diffusion models. *2023 IEEE Nuclear Science Symposium, Medical Imaging Conference and International Symposium on Room-Temperature Semiconductor Detectors (NSS MIC RTSD)*, pages 1–1.

Jiang, L., Mao, Y., Wang, X., Chen, X., and Li, C. (2023). Cola-diff: Conditional latent diffusion model for multi-modal MRI synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 398–408, Cham. Springer Nature Switzerland.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision – ECCV 2016*, pages 694–711, Cham. Springer International Publishing.

Jonsson, J., Karlsson, M., Karlsson, M., et al. (2010). Treatment planning using MRI data: an analysis of the dose calculation accuracy for different treatment regions. *Radiation Oncology*, 5(62).

Kläser, K., Markiewicz, P., Ranzini, M., Li, W., Modat, M., Hutton, B. F., Atkinson, D., Thielemans, K., Cardoso, M. J., and Ourselin, S. (2018). Deep boosted regression for MR to CT synthesis. In *Simulation and Synthesis in Medical Imaging*, pages 61–70, Cham. Springer International Publishing.

Largent, A., Barateau, A., Nunes, J.-C., Lafond, C., Greer, P. B., Dowling, J. A., Saint-Jalmes, H., Acosta, O., and de Crevoisier, R. (2019). Pseudo-CT generation for MRI-only radiation therapy treatment planning: Comparison among patch-based, atlas-based, and bulk density methods. *International Journal of Radiation Oncology, Biology, Physics*, 103(2):479–490.

Largent, A., Marage, L., Gicquiau, I., Nunes, J.-C., Reynaert, N., Castelli, J., Chajon, E., Acosta, O., Gambarota, G., de Crevoisier, R., and Saint-Jalmes, H. (2020). Head-and-neck MRI-only radiotherapy treatment planning: From acquisition in treatment position to pseudo-CT generation. *Cancer/Radiothérapie*, 24(4):288–297.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, X., Shang, K., Wang, G., and Butala, M. D. (2023a). DDMM-Synth: A denoising diffusion model for cross-modal medical image synthesis with sparse-view measurement embedding. *arXiv preprint arXiv:2303.15770*.

Li, Y., Xu, S., Chen, H., Sun, Y., Bian, J., Guo, S., Lu, Y., and Qi, Z. (2023b). CT synthesis from multi-sequence MRI using adaptive fusion network. *Computers in Biology and Medicine*, 157:106738.

Li, Y., Xu, S., Lu, Y., and Qi, Z. (2023c). CT synthesis from MRI with an improved multi-scale learning network. *Frontiers in Physics*, Volume 11 - 2023.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., and Guo, B. (2022). Swin transformer V2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002.

Lyu, Q. and Wang, G. (2022). Conversion between CT and MRI images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*.

Mei, S., Xia, Y., Fan, F., and Maier, A. (2025). GANeXt: A fully ConvNeXt-enhanced generative adversarial network for MRI- and CBCT-to-CT synthesis. *arXiv preprint arXiv:2512.19336*.

Meng, X., Gu, Y., Pan, Y., Wang, N., Xue, P., Lu, M., He, X., Zhan, Y., and Shen, D. (2022). A novel unified conditional score-based generative framework for multi-modal medical image completion. *arXiv preprint arXiv:2207.03430*.

Neppl, S., Landry, G., Kurz, C., Hansen, D., Hoyle, B., Stöcklein, S., Seidensticker, M., Weller, J., Belka, C., Parodi, K., and Kamp, F. (2019). Evaluation of proton and photon dose distributions recalculated on 2D and 3D Unet-generated pseudocts from T1-weighted MR head scans. *Acta Oncologica*, 58(10):1429–1434.

Nyúl, L. G. and Udupa, J. K. (1999). On standardizing the MR image intensity scale. *Magnetic Resonance in Medicine*, 42(6):1072–1081.

Oulbacha, R. and Kadoury, S. (2020). MRI to CT synthesis of the lumbar spine from a pseudo-3D cycle GAN. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1784–1787.

Owrangi, A. M., Greer, P. B., and Glide-Hurst, C. K. (2018). MRI-only treatment planning: benefits and challenges. *Physics in Medicine & Biology*, 63(5):05TR01.

Özbey, M. et al. (2023). Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539.

Pan, S., Abouei, E., Wynne, J., Chang, C.-W., Wang, T., Qiu, R. L. J., Li, Y., Peng, J., Roper, J., Patel, P., Yu, D. S., Mao, H., and Yang, X. (2024). Synthetic CT generation from MRI using 3D transformer-based denoising diffusion model. *Medical Physics*, 51(4):2538–2548.

Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. (2020). Contrastive learning for unpaired image-to-image translation. In *Computer Vision – ECCV 2020*, pages 319–345, Cham. Springer International Publishing.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2021). High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

Schmidt, M. A. and Payne, G. S. (2015). Radiotherapy planning using MRI. *Physics in Medicine and Biology*, 60(22):R323.

Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., and Fu, H. (2023). Transformers in medical imaging: A survey. *Medical Image Analysis*, 88:102802.

Smith, N. B. and Webb, A. (2010). *Frontmatter*, page i–vi. Cambridge Texts in Biomedical Engineering. Cambridge University Press.

Spadea, M. F., Maspero, M., Zaffino, P., and Seco, J. (2021). Deep learning based synthetic-CT generation in radiotherapy and PET: A review. *Medical Physics*, 48(11):6537–6566.

Thummerer, A., Galapon, A. J., Kurz, C., van der Bijl, E., Kamp, F., Landry, G., Terpstra, M., Maspero, M., Wahl, N., and Rogowski, V. (2024). Synthesizing computed tomography for radiotherapy challenge (synthrad2025).

Thummerer, A., van der Bijl, E., Galapon, A., Kamp, F., Savenije, M., Muijs, C., Aluwini, S., Steenbakkers, R., Beuel, S., Intven, M., Langendijk, J., Both, S., Corradini, S., Rogowski, V., Terpstra, M., Wahl, N., Kurz, C., Landry, G., and Maspero, M. (2025). Synthrad2025 grand challenge dataset: Generating synthetic cts for radiotherapy from head to abdomen. *Medical Physics*, 52.

Thummerer, A., van der Bijl, E., Galapon Jr, A., Verhoeff, J. J. C., Langendijk, J. A., Both, S., van den Berg, C. N. A. T., and Maspero, M. (2023). Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical Physics*, 50(7):4664–4674.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Villegas, F., Dal Bello, R., Alvarez-Andres, E., Dhont, J., Janssen, T., Milan, L., Robert, C., Salagean, G.-A.-M., Tejedor, N., Trnková, P., Fusella, M., Placidi, L., and Cusumano, D. (2024). Challenges and opportunities in the development and clinical implementation of artificial intelligence based synthetic computed tomography for magnetic resonance only radiotherapy. *Radiotherapy and Oncology*, 198:110387.

Wallimann, P., van Timmeren, J. E., Gabrys, H. S., Khodabakhshi, Z., Lapaeva, M., Dal Bello, R., Guckenberger, M., Andratschke, N., and Tanadini-Lang, S. (2025). N-Peaks: MRI intensity normalization based on normal tissue histogram peak intensities. *Physics in Medicine and Biology*, 70(24):24NT01.

Wasserthal, J., Breit, H.-C., Meyer, M. T., Pradella, M., Hinck, D., Sauter, A. W., Heye, T., Boll, D. T., Cyriac, J., Yang, S., Bach, M., and Segeroth, M. (2023). Totalsegmentator: Robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5).

Wieser, H.-P., Cisternas, E., Wahl, N., Ulrich, S., Stadler, A., Mescher, H., Müller, L.-R., Klinge, T., Gabryś, H., Burigo, L., Mairani, A., Ecker, S., Ackermann, B., Ellerbrock, M., Parodi, K., Jäkel, O., and Bangert, M. (2017). Development of the open-source dose calculation and optimization toolkit matRad. *Medical Physics*, 46(6):2556–2568.

Wolterink, J. M., Dinkla, A. M., Savenije, M. H. F., Seevinck, P. R., van den Berg, C. A. T., and Išgum, I. (2017). Deep MR to CT synthesis using unpaired data. In *Simulation and Synthesis in Medical Imaging*, pages 14–23, Cham. Springer International Publishing.

Xin, B., Sun, Z., Min, H., Belous, G., and Dowling, J. (2025). Team KoalAI: Ensembled ResUnet with masked anatomical perception loss. Algorithm description, SynthRAD2025 Challenge, Task 1.

Yan, S., Wang, C., Chen, W., and Lyu, J. (2022). Swin transformer-based GAN for multi-modal medical image translation. *Frontiers in Oncology*, Volume 12 - 2022.

Yi, X., Walia, E., and Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552.

Zeng, P., Zhou, L., Zu, C., Zeng, X., Jiao, Z., Wu, X., Zhou, J., Shen, D., and Wang, Y. (2022). 3D CVT-GAN: A 3D convolutional vision transformer-GAN for PET reconstruction. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 516–526, Cham. Springer Nature Switzerland.

Zhao, B., Cheng, T., Zhang, X., Wang, J., Zhu, H., Zhao, R., Li, D., Zhang, Z., and Yu, G. (2023). CT synthesis from MR in the pelvic area using residual transformer conditional GAN. *Computerized Medical Imaging and Graphics*, 103:102150.

Zhao, P., Pan, H., and Xia, S. (2021). MRI-Trans-GAN: 3D MRI cross-modality translation. In *2021 40th Chinese Control Conference (CCC)*, pages 7229–7234.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Zhu, L., Xue, Z., Jin, Z., Liu, X., He, J., Liu, Z., and Yu, L. (2023). Make-A-Volume: Leveraging latent diffusion models forcross-modality 3D brain MRI synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 592–601, Cham. Springer Nature Switzerland.