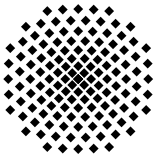

University of Stuttgart
Institute for Signal Processing and System Theory
Professor Dr.-Ing. B. Yang

Research Thesis S1449

Unsupervised Semantic Image Synthesis for Medical Imaging

Unüberwachte semantische Bildsynthese für medizinische Bilder

Author: Wenwu Tang

Date of work begin: 01, 07, 2023
Date of submission: 01, 03, 2024

Supervisor: M.Sc. George Basem Eskandar
Professor Dr.-Ing. B. Yang

Keywords: Deep learning, Generative Adversarial Networks, Medical Image Synthesis

Surgeries often necessitate the use of computed tomography (CT) for examining bony structures and magnetic resonance imaging (MR) for soft tissues. However, acquiring both CT and MR images is time-intensive, privacy sensitive and incurs significant costs. To address this challenge, we explore the potential of Generative Adversarial Networks (GANs) in generating synthetic labeled data. Our approach leverages a unique dataset comprising labeled CT scans with corresponding semantic labels and an unlabeled MR dataset without semantic annotations. The primary goal is to facilitate the translation of 2D CT semantic maps to 2D MR images, a task traditionally hampered by the requirement of supervised deep learning models for pairwise aligned training images-a rarity in the medical field. In response to these limitations, we introduce an innovative unsupervised method utilizing unpaired images to train our GAN model, termed Med-USIS. Through quantitative evaluations, Med-USIS has proven its efficacy in synthesizing MR images that closely approximate actual MR scans in terms of quality. Notably, it's close to conventional GAN models trained on paired MR and CT images, highlighting its advanced capabilities in image synthesis.

Contents

1. Introduction **1**
1.1. Motivation . . . 1
1.2. Outline . . . 3

2. Backgrounds and Related Works **5**
2.1. Preliminaries . . . 5
2.1.1. Generative Adversarial Networks (GANs) . . . 5
2.1.2. Domain Adaptation . . . 6
2.1.3. Unpaired Semantic Image Synthesis . . . 7
2.1.4. Medical Image Synthesis . . . 7
2.1.5. Frequency-based Deep Learning Scheme . . . 8
2.2. The USIS Baseline . . . 9
2.3. Problem Definition . . . 10

3. Proposed Method **13**
3.1. Med-USIS architecture . . . 13
3.1.1. Med-USIS Generator . . . 13
3.1.2. Med-USIS Discriminator . . . 16
3.1.3. U-Net based Segmentator . . . 16
3.1.4. Label Consistency Self-Supervision . . . 17
3.1.5. Mask Consistency Self-Supervision . . . 18
3.2. Pre-processing Methods . . . 19
3.2.1. Artifacts Removal . . . 20
3.2.2. Data Augmentation . . . 21
3.2.3. Normalisation, Resizing and One-hot Coding . . . 21

4. Experiments and Results **23**
4.1. Datasets . . . 23
4.1.1. AutoPET . . . 23
4.1.2. SynthRAD2023 . . . 24
4.2. Evaluation metrics . . . 25
4.2.1. Fréchet Inception Distance (FID) . . . 25
4.2.2. Structural Similarity Index (SSIM) . . . 25
4.2.3. Learned Perceptual Image Patch Similarity (LPIPS) . . . 26
4.2.4. Mean Intersection over Union (MIoU) . . . 26
4.2.5. Mean Absolute Error (MAE) . . . 27
4.2.6. Peak Signal-to-Noise Ratio (PSNR) . . . 27
4.2.7. Root Mean Square Error (RMSE) . . . 27
4.3. Experimental Setup . . . 28

4.4. Ablation Study 28
4.4.1. Ablation Study on the Generator Type 29
4.4.2. Ablation Study on the Mask Consistency Self-Supervision 31
4.4.3. Ablation Study on 3D Noise Input 33

5. Conclusion and Outlook **37**
5.1. Conclusion 37
5.2. Outlook 38

A. Additionally **39**
A.1. More generated images 39
A.2. More tables 45

List of Figures **47**

List of Tables **49**

Bibliography **51**

1. Introduction

1.1. Motivation

In medical domain, computed tomography (CT) and magnetic resonance imaging (MRI) are two essential and commonly utilized medical imaging technique. An example of CT, MR images from SynthRAD2023 dataset [1] is shown in figure 1.1. CT imaging, excels in providing electron density information. However, they typically lack the excellent soft-tissue contrast provided by MR imaging. Although, MR imaging is praised for its ability to provide high-quality soft-tissue contrast, they do not directly offer information about electron density. To address the complementary strengths and weaknesses of MR and CT imaging, it is common in clinical practice to acquire both types of images for a comprehensive understanding of the patient's anatomy. However, obtaining both MR and CT images involves significant time, cost and the necessity for precise MR-CT registrations to fully leverage the benefits of the combined dataset. If the acquisition of MR scans could be omitted, it would alleviate these challenges.

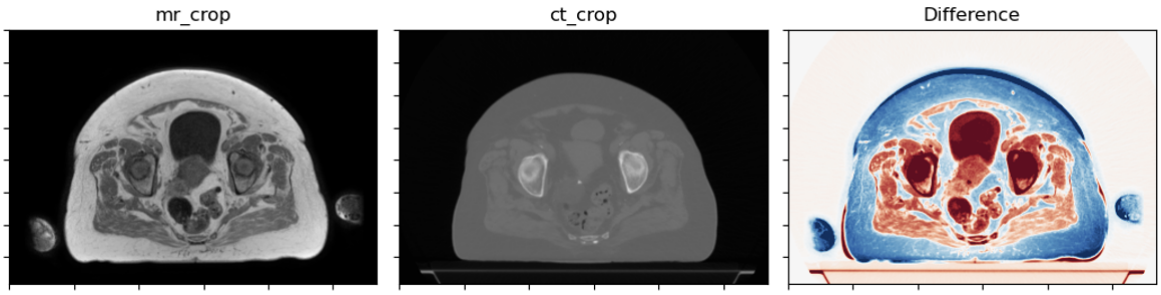

Figure 1.1.: A pair of corresponding MR (left) and CT (right) images and their difference

The methodology of converting CT labels (semantic labels or segmentation maps) to MR images holds paramount importance, particularly from the perspective of patient privacy preservation. This advanced approach leverages high-level, structured information extracted from original CT images, primarily focusing on labels that represent critical anatomical structures or regions. Unlike the direct conversion of CT images to MR images, which involves processing raw pixel data potentially laden with patient-specific details, this label-based method significantly reduces the risk of revealing sensitive personal information. The utilization of CT labels, as opposed to complete image data, aligns seamlessly with stringent privacy regulations such as HIPAA [2], ensuring that medical data sharing and utilization uphold the highest standards of privacy protection. By stripping away directly identifiable personal information and focusing on medically relevant, de-identified data, this approach not only complies with legal and ethical standards but also minimizes the privacy risks associated with data exposure during sharing or publication. Moreover, the transition from CT labels

to MR images offers a strategic advantage in the collaborative and interdisciplinary nature of medical research. It facilitates the sharing of crucial medical insights among professionals and researchers without compromising the sanctity of patient confidentiality. This balance between information sharing and privacy preservation is crucial in fostering innovation and advancing medical science while maintaining public trust and adherence to ethical standards.

Due to the big advantage of medical image synthesis, there has been a growing scholarly focus on deep learning-based methods for medical image synthesis [3], driven by their privacy preservation, exceptional efficacy in precise image mapping and notable computational efficiency compared to traditional models [4]. We also argue that, with guidance of semantic maps, model can improve the quality of generated MR images comparing with direct CT-MR translation [5]. CT semantic maps are easy to obtain, which can be manually labeled by professionals or automatically generated by segmentation networks. As research on CT segmentation models advances, existing models such as TotalSegmentator [6] based on nnU-Net [7] offers pre-trained models that facilitate rapid and accurate acquisition of semantic maps from CT images. However, a predominant limitation within the current landscape of deep learning synthesis approaches lies in their stringent reliance on large scales of paired data. This constraint becomes particularly pronounced in scenarios where acquiring meticulously registered image pairs is impractical or unfeasible. If models can be trained effectively using unpaired data, the constraints related to data availability are considerably alleviated. This paradigm shift could lead to more robust models that are capable of learning from a diverse array of datasets, not limited by the prerequisite of paired samples. Consequently, this approach holds the potential to revolutionize the field by mitigating one of the fundamental bottlenecks - the need for large, accurately annotated, and paired datasets. It opens up new possibilities for leveraging extensive, previously untapped datasets, thereby accelerating advancements in medical imaging analysis and other domains reliant on large-scale data (figure 1.2).

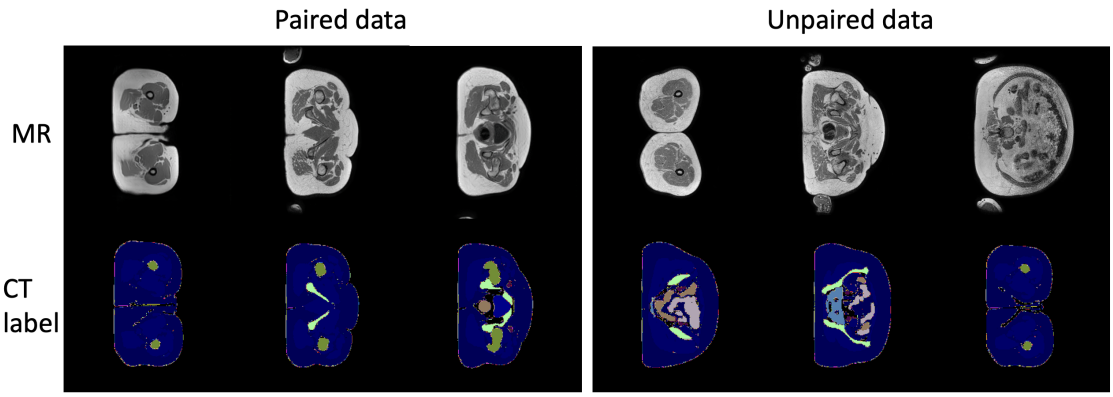

Figure 1.2.: Example of paired data and unpaired data

A notable improvement in this domain is the development of the Med-USIS model, which represents a paradigm shift in the synthesis of medical images. Uniquely, this model utilizes unpaired data for training, marking a significant departure from the traditional dependency on meticulously curated paired datasets. This innovative approach not only broadens the scope of image synthesis methodologies but also presents a practical solution for scenarios where obtaining perfectly matched data pairs is challenging or resource-intensive. The Med-USIS

model opens up new possibilities for medical image analysis and synthesis, particularly in areas where data scarcity or privacy concerns limit the availability of paired datasets. By leveraging unpaired data, Med-USIS demonstrates the potential to enhance medical training, and generation procedures.

1.2. Outline

First, the background and related works are introduced in Section 2. This section includes preliminaries, which consist of an overview of Generative adversarial networks, Semantic image synthesis and Frequency-based Deep Learning Scheme, our baseline model USIS [8] and the definition of the problem to be solved. Section 3 introduces our proposed method: the Med-USIS model architecture, loss functions and pre-processing methods. In Section 4, experiments are designed to compare the performance of different generator, verify the validity of mask consistency loss and investigate how noise input impact the generated images. Section 5 concludes our work and proposes the potential direction for future research.

2. Backgrounds and Related Works

2.1. Preliminaries

2.1.1. Generative Adversarial Networks (GANs)

Generative Adversarial Networks (GANs)[9] represent a groundbreaking paradigm in the field of deep learning, introduced by Ian Goodfellow and collaborators in 2014. GANs have since become a cornerstone in various applications, showcasing their remarkable ability to generate realistic data through an innovative adversarial training approach.

The main architecture of GAN is a dual-network framework comprising a Generator and a Discriminator. The Generator generates synthetic samples from random noise, aiming to replicate real data and fool the Discriminator, while the Discriminator assesses and distinguishes between real and generated samples. Through iterative adversarial training, the Generator refines its ability to produce increasingly realistic data, while the Discriminator enhances its capacity to discern between real and synthetic samples. In this equation 2.1, the first term represents the expected value of the logarithmic output of the discriminator for real data x, which is sampled from the true data distribution $p_{\text{data}}(x)$. The second term calculates the expected value of the logarithmic output of one minus the discriminator's output for fake data. This fake data is produced by the generator using input noise z, which is sampled from the noise distribution $p_z(z)$. The generator's goal is to minimize this objective function, while the discriminator aims to maximize it, creating a competitive game that ideally leads to the generation of realistic synthetic data. The structure is shown in figure 2.1

$$\min_G \max_D L(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.1}$$

The unique adversarial dynamic between the two networks empowers GANs to generate data that convincingly mimics real-world distributions. This powerful capability has propelled GANs into diverse domains, including image generation, image-to-image translation, style transfer, and beyond.

The Conditional Generative Adversarial Network (CGAN) [10] represents an extension of the traditional GAN framework by integrating a conditional variable. In CGANs, both the generator (G) and the discriminator (D) receive extra information as input, which could be an image, a label, or any other auxiliary data. This conditioning directs the CGAN to produce output that aligns with particular characteristics, for instance, generating images of a certain class or emulating attributes within a given dataset. The objective function for a CGAN is formulated to reflect this conditional generation process as follows:

$$\min_G \max_D L(G, D) = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x|y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z|y)))] \tag{2.2}$$

Here, the term $\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x|y)]$ represents the expectation of the discriminator's log-probability that real data x, conditioned on y, is authentic. Conversely, $\mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z|y)))]$ denotes the expectation of the discriminator's log-probability that a fake instance $G(z|y)$, generated from noise z and conditioned on y, is not authentic. The generator seeks to minimize this function, thereby fooling the discriminator into thinking the generated instances are real, while the discriminator aims to maximize it, accurately distinguishing between real and generated instances.

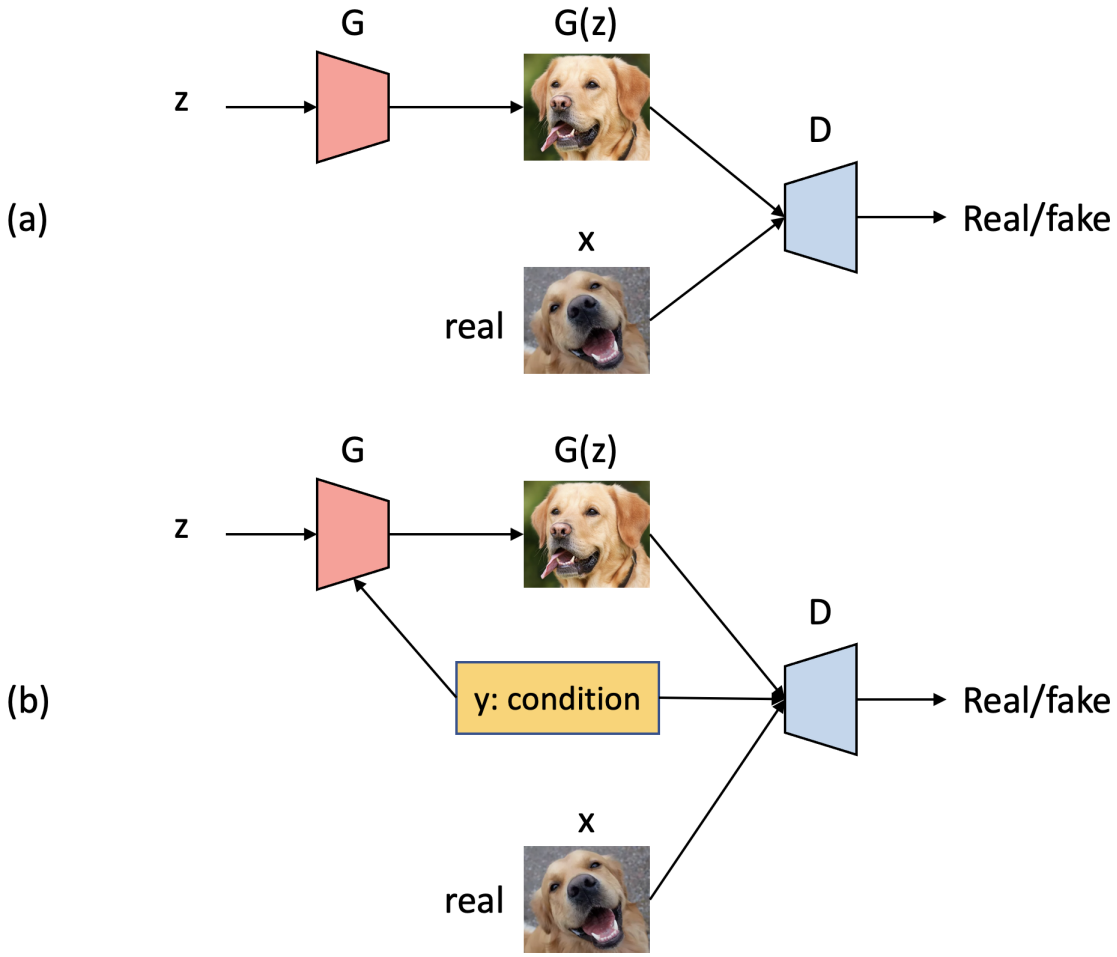

Figure 2.1.: (a) GAN architecture (b) Conditional GAN (CGAN) architecture

2.1.2. Domain Adaptation

Domain adaptation is a machine learning paradigm designed to address the challenge of model generalization when confronted with variations in the distribution of data between different domains. In various real-world scenarios, models trained on a source domain may encounter a drop in performance when applied to a target domain with different characteristics [11]. Domain adaptation methods aim to enhance the adaptability of models by mitigating the distributional shift between the source and target domains, thereby improving their effectiveness in diverse and dynamic environments. This field is particularly crucial in applications where labeled data in the target domain is scarce or expensive to acquire, as domain

adaptation techniques facilitate the transfer of knowledge from a labeled source domain to an unlabeled or sparsely labeled target domain [12].

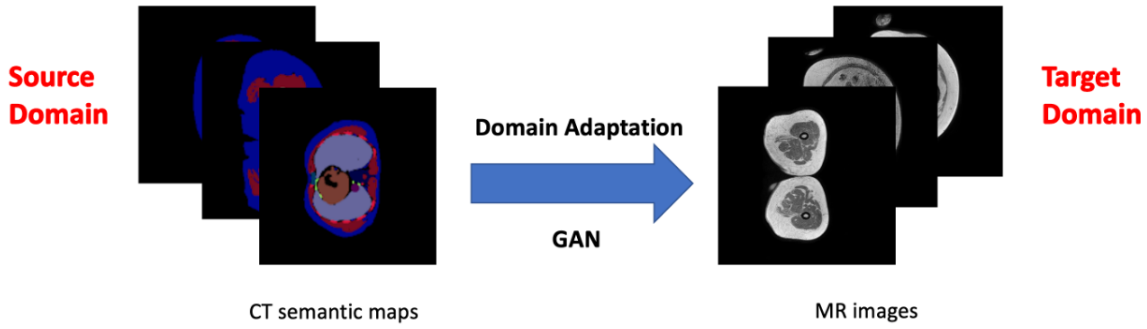

Figure 2.2.: Our task described as a domain adaptation problem

This approach reduce the domain gap between the CT labels (source domain) and the T1 weighted MR images (target domain). The challenge exists in the distinct data distributions and imaging characteristics of CT and MR modalities. Therefore, the primary objective is to devise a learning model capable of effectively adapting the semantic and structural information in the CT labels to generate corresponding MR images that are not only anatomically accurate but also exhibit the imaging properties characteristic of genuine MR scans. This domain adaptation not only necessitates the preservation of critical medical information during the transformation but also demands the synthesized MR images to be of diagnostic quality, ensuring their utility in subsequent medical analysis and decision-making processes.

2.1.3. Unpaired Semantic Image Synthesis

Due to the high cost of collecting paired medical image, we utilize unpaired paradigm [13, 14, 15]. This technique stands out in its ability to generate semantically coherent and visually plausible images without the prerequisite of paired training datasets . Traditional image synthesis methods heavily relied on paired data, where direct correspondence between input and output images was a necessity. However, the advent of unpaired methods has revolutionized this approach, offering a new paradigm for image generation. Unlike conventional methods that require pixel-level correspondence, this approach understands the underlying semantic context, enabling the transformation of images across diverse domains without direct pairing. For instance, transforming a daytime landscape into a nighttime scene or altering weather conditions within an image are feats achievable through this technique.

2.1.4. Medical Image Synthesis

In the field of medical informatics and computer vision, medical image synthesis plays a pivotal role, tasked with generating highly realistic images for training, research, and clinical diagnosis. This synthesis process, however, faces multifaceted challenges and key considerations unique to medical applications. Paramount among these is the need for exceptional fidelity and accuracy, as synthesized images are often integral to diagnostic processes where even minor inaccuracies can have significant patient care implications. The complexity of

human anatomy, with its vast array of variations, further complicates synthesis. Models must adeptly represent diverse anatomical structures, pathologies, and patient demographics, addressing the inherent variability in human physiology. Compounding these challenges are the constraints of data availability, often limited by privacy concerns and the rarity of certain medical conditions, necessitating compliance with stringent regulations like HIPAA [2].

Medical image synthesis also demands versatility across various imaging modalities, each with distinct characteristics. A comprehensive understanding of how these modalities-ranging from MRI and CT to X-ray and Ultrasound-correlate with different tissue types and pathologies is essential. The synthesis process must strike a delicate balance between realism and interpretability, especially in scenarios where enhanced or exaggerated features are vital for educational or diagnostic objectives.

Furthermore, the challenge of unpaired image synthesis is prominent, especially in the context of rare medical conditions where paired images (e.g., pre- and post-treatment) are scarce. Adapting to the continuously evolving landscape of medical technology and diagnostic standards is also crucial, requiring synthesis models to be both flexible and up-to-date.

Addressing these complexities in medical image synthesis calls for an interdisciplinary approach, combining expertise from medical science, computer vision, machine learning, and ethical considerations. This collaborative effort is essential in advancing the field and developing robust, effective synthesis models that meet the high standards required in medical applications.

2.1.5. Frequency-based Deep Learning Scheme

Several works [16, 17, 18] found the notable absence of high-frequency content in the discriminator component of standard Generative Adversarial Networks (GANs). This caused by the downsampling layers commonly employed in GAN architectures. Such a configuration inadvertently leads to a spectrum discrepancy: the generator, lacking adequate incentive from the discriminator, fails to learn and replicate the high-frequency content of data. This results in a marked divergence between the spectral characteristics of generated images and real images. In order to preserve the high-frequency content, it's necessary to get the representation of images in frequency domain e.g., 2D-Fourier transform or Haar wavelet transform. [19] incorporate a wavelet packet transform (WPT) module, which can capture texture details at multiple scales within the frequency space so that enhance the visual fidelity of the generated images. SSD-GAN[16] proposed a novel enhancement designed to mitigate spectral information loss within the discriminator, embedding a frequency-aware classifier into the discriminator, which is able to assess the 'realness' of inputs across both spatial and spectral domains SWGAN[20] made the first attempt to fully utilize frequency-based approach to generative model, they distinctively integrated wavelets within both the generator and discriminator architectures, ensuring a frequency-aware latent representation at every stage of the generation process. The experiments confirm that synthesizing content in the wavelet domain leads to higher-quality images, particularly by enriching the realism of high-frequency content. We choose wavelet representation instead of Fourier representation, since it can represent the images in both space and spectral domains.

Wavelets allow for the analysis of images at multiple scales or resolutions. This is crucial in medical imaging, where different levels of detail may be required to accurately diagnose

a range of conditions. For instance, finer details may be needed to analyze microcalcifications in mammograms, while broader features are essential in MRI brain scans. Wavelets are particularly effective in preserving edges and textures, which are vital in medical images. Accurate representation of edges and textures is essential for the diagnosis and analysis of various medical conditions, as these features often hold significant diagnostic information. Wavelets can be used to enhance specific features of an image, such as anomalies or regions of interest, making them more visible for analysis. This is crucial in detecting subtle pathological changes that might be missed in standard imaging processes.

In summary, the wavelet approach offers a flexible, robust, and efficient method for medical image synthesis, addressing key challenges such as detail preservation, noise reduction, and multi-resolution analysis, which are essential for the accurate and effective use of medical images in clinical practice and research.

2.2. The USIS Baseline

USIS, or Unsupervised Semantic Image Synthesis [8] is a powerful model to synthesize multimodal realistic images from labels without the use of paired data. It involves a unique and effective learning scheme that amalgamates the fragmented advantages of cycle losses and relationship preservation constraints.

USIS utilize SPADE [21] Generator $\mathcal{G}$, StyleGAN2 [22] discriminator $\mathcal{D}$ and a U-Net [23] based segmentation network $\mathcal{S}$.

The generator implements spatially-adaptive normalization as know as SPADE, which is designed to enhance the synthesis of photorealistic images from semantic layouts. Traditional methods, which process semantic layouts through deep networks with convolutional, normalization, and activation layers, often lead to the dilution of semantic information. Our method preserves this crucial information by adaptively modulating normalization layer activations with input layouts, utilizing a learned spatial transformation. This technique significantly improves visual realism and alignment with input layouts, setting a new standard in image synthesis.

The discriminator features a wavelet-based encoder and a decoder for reconstructing real images. This configuration, particularly the self-supervised reconstruction loss in the decoder, is designed to prevent encoder overfitting on limited wavelet coefficients. This methodology has been tested across three challenging datasets, setting a new benchmark for unpaired SIS. It achieved a landmark in the field of unpaired image generation, surpassing the current state-of-the-art by more than double in performance on the ADE20K [24] and COCO-Stuff [25] datasets.The results are compelling, with the generated images exhibiting enhanced diversity, quality, and multimodality, marking a notable advancement in the field of Semantic Image Synthesis.

In the described framework, the generator $\mathcal{G}$ is responsible for generating RGB images $\mathbf{x}$ from a given semantic map $\mathbf{m}$, which is one-hot encoded. Simultaneously, the discriminator $\mathcal{D}$ determinates whether the generated images are real or fake. A segmentation network $\mathcal{S}$ complements this process by attempting to reconstruct the original semantic map $\mathbf{m}$ from the generated images. It's worth noting that $\mathcal{S}$ exclusively interacts with the generated images, contrasting with $\mathcal{D}$, which evaluates both real and synthesized images. This architecture

fosters a competitive relationship between $\mathcal{D}$ and G, with the goal of refining $\mathcal{G}$ to create highly photorealistic images. Concurrently, $\mathcal{S}$ and $\mathcal{G}$ collaborate, leveraging a class-balanced self-supervised segmentation loss $\mathcal{L}_{seg}$ based on the input mask. This cooperative dynamic ensures that the generated images not only exhibit high visual quality but also maintain semantic consistency with the original input, aligning closely with the provided semantic maps.

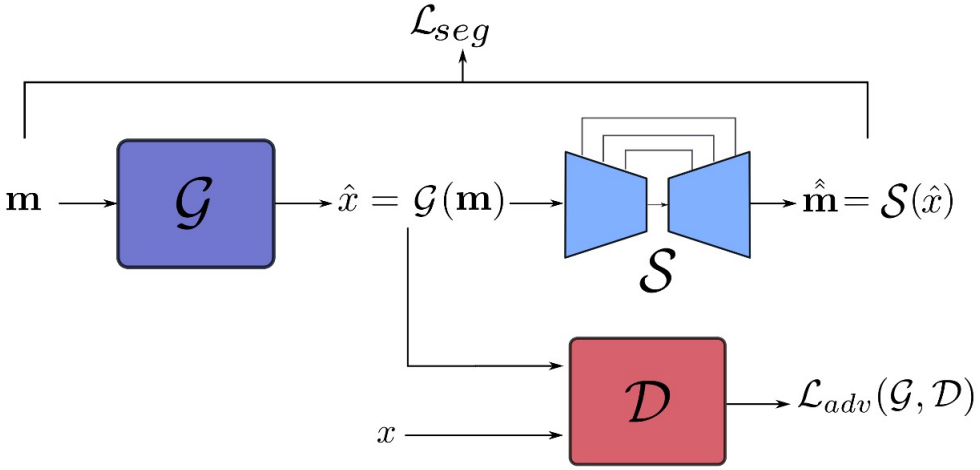

Figure 2.3.: USIS model architecture [8]

USIS doesn't require paired datasets (i.e., strictly corresponding CT MR images of the same patient) for training. This is a significant advantage in the medical field, where acquiring such paired datasets can be challenging due to ethical, legal, and privacy constraints. Medical images exhibit a wide range of variability due to differences in patient anatomy, disease manifestations, and imaging modalities. USIS can learn from unpaired datasets to understand and replicate this variability, making it better suited for synthesizing realistic medical images that can accommodate diverse cases. USIS's flexibility allows it to adapt to different modalities and synthesize images across these diverse formats without needing exact paired examples. USIS can effectively learn from unstructured or unlabeled data, which is a common scenario in medical databases. This ability enables it to extract meaningful patterns and features from a wide array of medical images, enhancing the synthesis process. USIS's ability to learn from unstructured and unpaired data, coupled with its adaptability and cost-effectiveness, makes it highly suitable for addressing the unique challenges of medical image synthesis.

2.3. Problem Definition

In this section, we delineate the task of semantic medical image synthesis and we will expound upon the role of our model in tackling this challenging objective in next section. Our goal is the unsupervised synthesis of RGB MR images based on CT labels. Labels can be either manually annotated or even automatically generated by segmentation tools. The ensuing section provides a comprehensive discussion of our task.

We possess a CT dataset, consisting of CT scans paired with their respective semantic masks, and an unlabeled dataset of MRI includes MRI scans without associated semantic labels. We

seek to learn a mapping from semantic CT mask $\boldsymbol{m_{CT}}$ with C classes to corresponding MR image $\mathbf{x}$ in a unsupervised way. The mask $\boldsymbol{m_{CT}}$ is one-hot coded, which has C channels, and CT images are normalized between 0 and 1.

The challenge at hand can be distilled into two fundamental tasks for effective Unpaired Semantic Image Synthesis in the medical imaging domain. Firstly, the task of achieving class appearance alignment is crucial. This requires that specific organs and tissues in the generated images possess the correct texture, conforming to their expected visual properties. The discriminator plays a vital role in this process, tasked with discerning the association between semantic classes within the segmentation map and their corresponding textures in the actual image. Secondly, it is imperative to ensure shape consistency in every organ and tissue depicted in the generated images. This demands that the generator not only reproduces the visual appearance but also upholds the content and geometrical structure inherent in the segmentation map. Such fidelity is particularly crucial for the nuanced representation of minor classes, which may be easily overlooked but are critical in medical imaging. Addressing these challenges requires a tailored approach, one that is acutely sensitive to the unique requirements of medical imaging. It involves a fine balance between the accuracy of texture representation and the precision of structural alignment, ensuring that the synthesized images are both visually coherent and medically relevant. This paper aims to explore and optimize methodologies to meet these intricate objectives, thereby advancing the field of medical image synthesis through Unpaired Semantic Image Synthesis.

3. Proposed Method

In this section, we utilize Med-USIS architecture, shown in figure 3.1 for unpaired semantic medical image synthesis, This architecture is heavily based on the USIS model. We adapt USIS model to application in medical domain. First, we present the generator and discriminators in detail, then we introduce the proposed loss components. To the best of our knowledge, it is the first work to generate MR images from CT labels with unpaired training data in a unsupervised paradigm.

3.1. Med-USIS architecture

The network comprise four components: A wavelet generator $\mathcal{G}$, A wavelet discriminator $\mathcal{D}$ and a U-Net based segmentation network $\mathcal{S}$ and two mask extractors that extract the mask from input label map and generated MR image, the mask loss minimizes the L_2 distance between the the two extracted masks, in order to penalize the shape difference and ensure accurate generation. The parameter count of model components is shown in table 3.1

3.1.1. Med-USIS Generator

In our approach, we generate a range of diverse outputs directly from input noise, bypassing the need for complex mechanisms. This is achieved by creating a noise tensor with 64 channels and spatial dimensions identical to the label map ($H \times W$). By channel-wise concatenating this noise tensor with the label map, we produce a 3D tensor that serves as the initial input to the generator. Furthermore, this combined tensor is utilized as a conditioning element at every spatially-adaptive normalization layer throughout the network. As a result, the intermediate feature maps are influenced by the semantic labels as well as the input noise, ensuring that the generated outputs are varied and contingent on the input noise pattern.

The provided equation 3.1 represents the loss function for the generator $\mathcal{L}_{Gen}$ in a Generative Adversarial Network, specifically in the context of semantic segmentation and generation. This composite loss function combines the segmentation loss $\mathcal{L}_{seg}$, the adversarial loss $\mathcal{L}_{advG}$ and mask consistency loss $\mathcal{L}_{mask}$ to train the generator. λ_{seg} is the weighting factor for the segmentation loss. $\mathcal{L}_{seg}(m, \mathcal{S}(\mathcal{G}(m)))$ is the segmentation loss, measuring the discrepancy

Table 3.1.: Parameter count of model components

Name	OASIS Generator	Wavelet Generator	Wavelet Discriminator	U-Net
#param	71.1M	55.8M	26.2M	22.3M

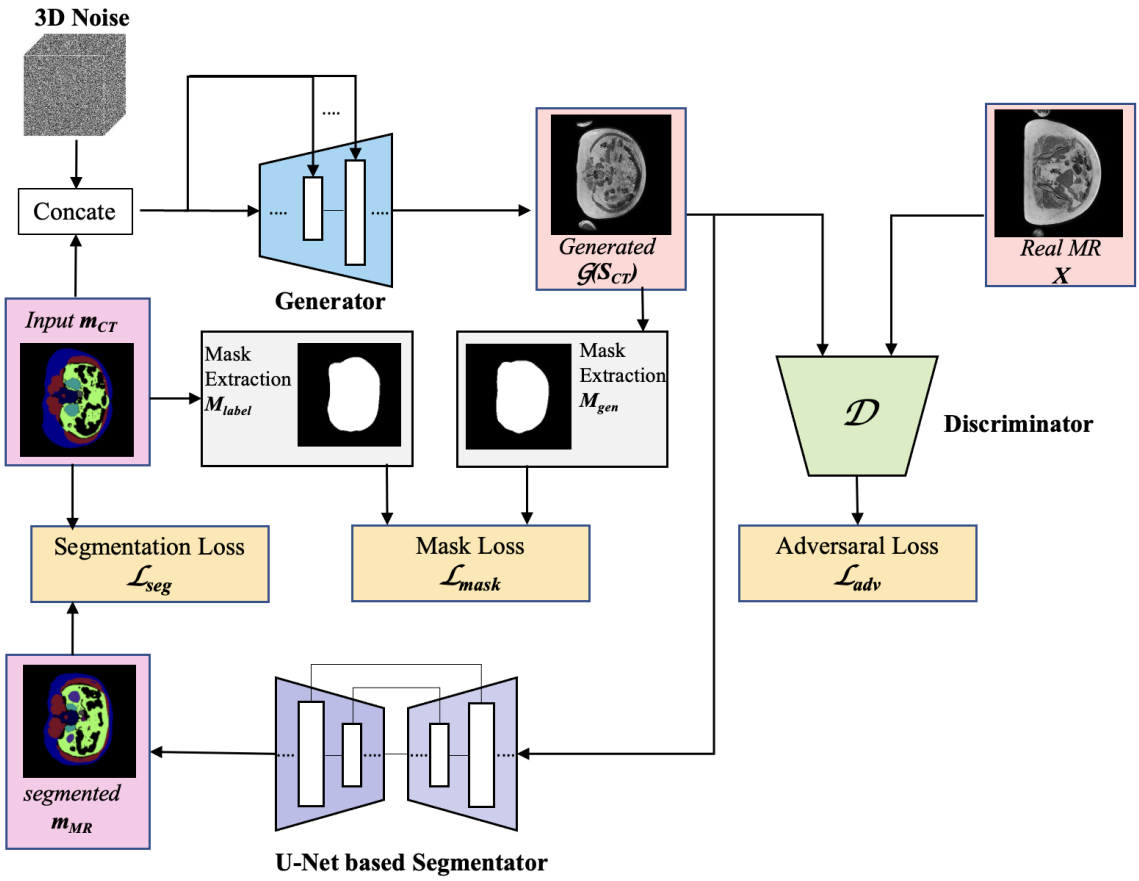

Figure 3.1.: Overview of our proposed Med-USIS model

between the true segmentation mask m and the predicted mask $\mathcal{S}(\mathcal{G}(m))$. $\mathcal{L}_{advG}(D(G(m)))$ is the adversarial loss for the generator, quantifying the generator's success in fooling the discriminator.

$$\mathcal{L}_{Gen} = \lambda_{seg}\mathcal{L}_{seg}(m, \mathcal{S}(\mathcal{G}(m))) + \mathcal{L}_{advG}(\mathcal{D}(\mathcal{G}(m))) + \lambda_{mask}\mathcal{L}_{mask}(m, \mathcal{G}(m)) \tag{3.1}$$

This loss function ensures that the generator not only produces images that are indistinguishable from real images by the discriminator but also accurately represents the semantic segmentation of the scene.

We implement OASIS generator and Wavelet generator. Both of them are based on SPADE [21] and Residual connection [26]. In SPADE, the mask undergoes a transformation into an embedding space and is further processed through convolution to generate modulation parameters γ and β as shown in figure 3.2. What sets SPADE apart from previous conditional normalization methods is that γ and β are not vectors, but rather tensors that retain spatial information. These parameters, γ and β, are then used in a spatially-aware manner: they are multiplied and added to the normalized activation for each element. This approach ensures that the specific details and structure in the mask directly influence the modulation at every location in the feature map.

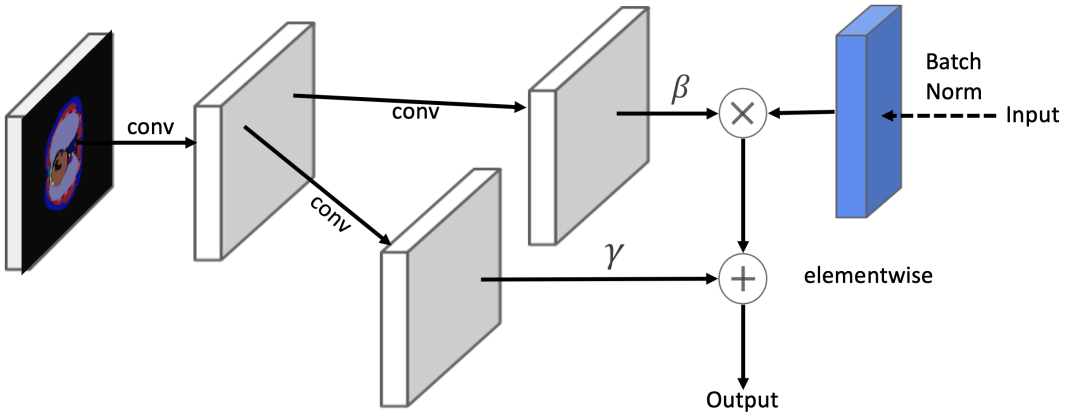

Figure 3.2.: SPADE structure [21]

OASIS Generator

OASIS, or You Only Need Adversarial Supervision For Semantic Image Synthesis [27], represents a paradigm shift in generative adversarial networks designed for semantic image synthesis. OASIS stands out by discarding the conventional reliance on VGG-based perceptual loss [28], which has been a staple in improving the image quality of GAN outputs. The OASIS generator leverages 6 Resnet Blocks integrated with Spatially-Adaptive Normalization (SPADE) to ensure the generated images preserve the semantic layout provided by the input segmentation masks. The initial layer takes a semantic map as input, optionally augmented with a noise vector and edge information if specified. This combination allows the network to introduce variations and capture finer details, enhancing the diversity and quality of the generated images. Each Resnet Blocks equips with 2 SPADE normalization and Residual connection. SPADE is designed to retain the semantic information of the input map throughout the network, ensuring that each location in the output image gets properly conditioned on the corresponding location in the input semantic map. This mechanism is particularly beneficial for generating images that are not only realistic but also semantically coherent. The network progressively upsamples the feature maps to the desired resolution. In the process, it employs skip connections and optional progressive growing techniques to stabilize the training and improve the quality of the generated images.

Wavelet Generator

Wavelet generator is a unique neural network designed for creating detailed and realistic images. It starts by mixing input information like the shape of objects (semantic map) and random noise to generate diverse results. The network uses special building blocks called Residual Wavelet Blocks. These blocks harness the power of wavelet transformations to meticulously analyze and enhance details at multiple levels, while also incorporating SPADE to ensure that the semantic integrity of the input is preserved throughout the process. As the network processes the input, it gradually upscales the image, ensuring that the finer details are not lost. In the end, the network combines all the learned information to produce a final image that is both detailed and closely matches the input shape. This approach allows our network to generate high-quality images that are rich in detail and visually appealing.

3.1.2. Med-USIS Discriminator

The Wavelet discriminator network begins with a Haar wavelet transformation of the input, ensuring that the network operates on multi-resolution representations of the images. This approach allows the network to analyze images at various scales and capture both coarse and fine details effectively. A series of convolutional blocks process the feature maps. Each block consists of convolutional layers that transform the feature map from a higher resolution to a lower one. The number of channels in each block is predefined based on the image's resolution, ensuring that the network can capture relevant features at each scale.

This loss function 3.2 is instrumental in training the discriminator to accurately distinguish between real and generated images, a critical aspect of the adversarial training process in GANs. $\mathcal{L}_{advD}$ represents the adversarial loss for the discriminator. $\mathcal{D}(x)$ is the discriminator's output for the real images, where the discriminator aims to output values close to 1. $\mathcal{D}(\mathcal{G}(m))$ is the discriminator's output for the generated images, where the discriminator aims to output values close to 0.

$$\mathcal{L}_{Dis} = \mathcal{L}_{advD}(\mathcal{D}(x), \mathcal{D}(\mathcal{G}(m))) \tag{3.2}$$

3.1.3. U-Net based Segmentator

The presented figure 3.3 illustrates our U-Net based architecture, a prevalent design for tasks requiring precise spatial information reconstruction such as image segmentation and image-to-image translation. It's not just a segmentator, but also a discriminator [29]. The self-supervised segmentation loss that come from the segmentator can be seen as a relationship preservation contraint. The network commences with an input of a 3-channel image that come from generator. This input is initially processed through a 3x3 convolutional layer, followed by a series of residual blocks that expand the feature channels to 128. Subsequent layers involve a downsampling step via 2x2 average pooling, reducing the spatial dimensions by half while progressively increasing the channel depth up to 512. The decoder component of the network begins with an upsampling operation that doubles the spatial resolution of the feature maps. These upsampled maps are then concatenated with their corresponding feature maps from the encoder through skip connections, which facilitate the retention of spatial context lost during downsampling. After merging, the features pass through additional residual blocks. This process repeats, with each upsampling step followed by concatenation and residual processing until the feature maps are reduced to 64 channels. Finally, the network concludes with a 1x1 convolutional layer that maps the 64-channel feature representation back to the desired (C+1) channel output. C is the number of classes. Beyond the C semantic classes derived from the label maps, all pixels within the synthetic images are classified into an additional, separate category. and we propose (C+1) cross-entropy loss for training. This architecture ensures that the network leverages context information captured during encoding to generate detailed and accurate outputs, making it highly suitable for applications requiring detailed feature localization and preservation.

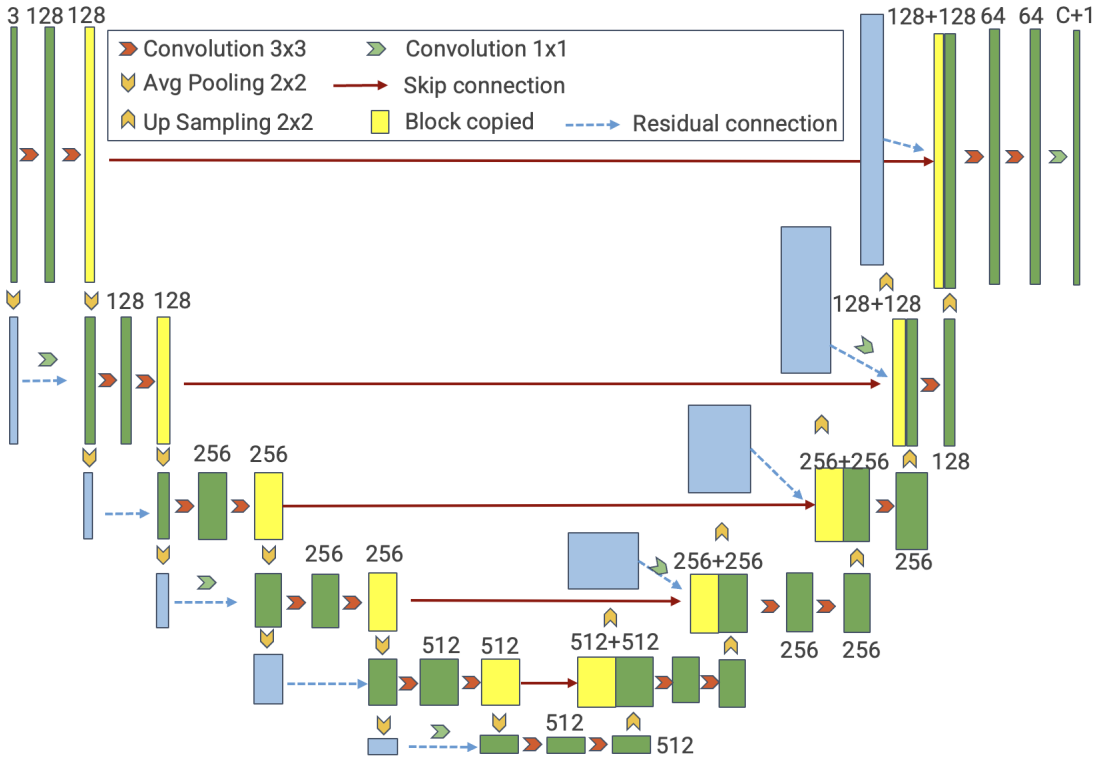

Figure 3.3.: Structure of the U-Net based Segmentator

3.1.4. Label Consistency Self-Supervision

This loss function $\mathcal{L}_{seg}$ is used to train the U-Net segmentation network by comparing the predicted segmentation map against the ground truth with cross-entropy loss, with class weights addressing the imbalance in the dataset. The weighted cross-entropy loss function for semantic segmentation in the context of Generative Adversarial Networks is defined as:

$$\mathcal{L}_{seg} = -\mathbb{E}_m \left[\sum_{c=1}^{C} \alpha_c \sum_{i=1}^{H} \sum_{j=1}^{W} m_{c,i,j} \log(\mathcal{S}(\mathcal{G}(m))_{c,i,j}) \right] \tag{3.3}$$

In the provided function 3.3, C represents the number of classes involved in the segmentation task, while H and W denote the height and width of the image, respectively. The term α_c refers to the class-specific weight, introduced to mitigate issues stemming from class imbalance. The notation $m_{c,i,j}$ corresponds to the ground truth label for class c at the specific pixel location (i, j), serving as a reference for what the model's prediction should aim to replicate. Lastly, $\mathcal{S}(\mathcal{G}(m))_{c,i,j}$ denotes the predicted probability that the model assigns to the pixel at location (i, j) for belonging to class c, reflecting the model's assessment of the image content at that particular pixel.

Class-balancing weights α_c are crucial in addressing the issue of class imbalance in our tasks, since The imbalance in our datasets is significant as shown in table A.3 . These weights are calculated to be inversely proportional to the per-pixel class-frequency, ensuring that less frequent classes are given higher importance during the training process. The weight for each class c can be mathematically formulated as:

$$\alpha_c = \frac{H \times W}{\sum_{i,j}^{H \times W} \mathbb{E}_m \left[1[m_{c,i,j}] \right]} \tag{3.4}$$

The indicator function, denoted as $1[m_{c,i,j}]$, is defined for a pixel at position (i, j) and class c as follows:

$$1[m_{c,i,j}] = \begin{cases} 1, & \text{if the pixel at position } (i, j) \text{ belongs to class } c \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

By employing these weights, the impact of class imbalance is mitigated, allowing the model to train in a more balanced and effective manner.

The loss function for the U-Net architecture in the context of semantic segmentation is defined as:

$$\mathcal{L}_{UNet} = \mathcal{L}_{seg}(m, \mathcal{S}(\mathcal{G}(m))) \tag{3.6}$$

This loss function ensures that the U-Net model is trained to generate segmentation maps that closely match the ground truth, thereby facilitating accurate and effective segmentation. To illustrate the performance of the U-Net model, we process the output of $\mathcal{S}$ by applying colorization to the segmented regions. It is observed that the output is rather noisy, particularly in the background regions, indicating that the U-Net struggles with precise class classification. Misalignments between the predicted and actual classes are evident, though the general shapes of the objects are well retained in the segmentation. Managing the classification of 31 distinct classes simultaneously poses a substantial challenge for a standard U-Net model, which may contribute to the observed difficulties in achieving clear and accurate class separation. We take the model Wavelet generator with mask loss as an example. We exemplify the performance of the model using the Wavelet generator enhanced with mask loss. The images depicted in figure 3.4 showcase the outputs generated by this model. To evaluate the segmentation quality, we employ a two-step process: firstly, we input the MR images synthesized by the model into a U-Net to obtain Segmentation 1. Subsequently, we feed the actual MR images into the same U-Net to produce Segmentation 2.

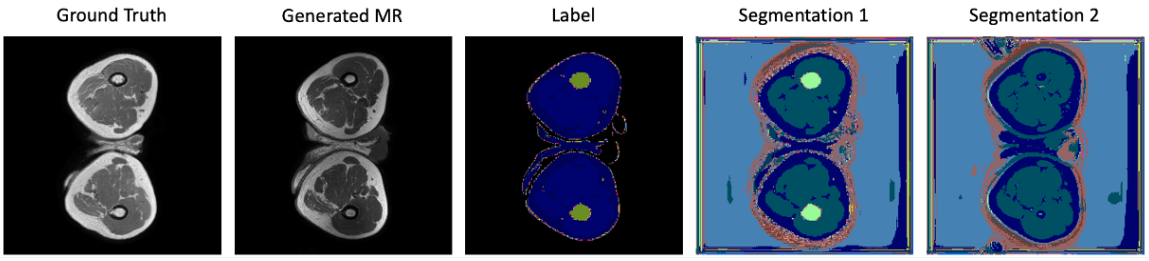

Figure 3.4.: Outputs of the U-Net

3.1.5. Mask Consistency Self-Supervision

We propose the mask consistency to enforce this critical constraint: in an ideal synthesis result, the shape and external contours of generated MR image and CT label of the same pa-

tient should perfectly correspond to each other. This alignment ensures that the synthesized MR images are not only accurate but also clinically relevant, maintaining a high degree of fidelity to the anatomical structures represented in the CT labels. Instead of using neural network to learn the mask, we utilize basic image processing methods such as Gaussian blurring and binary Opening operations to extract mask of both generated MR images and CT labels. First, we implement Gaussian blur as an initial step to effectively reduces noise and simplifies the features of an image by smoothing out small details, which helps in focusing on larger, more significant structures, rather than getting distracted. In binary operation stage, erosion removes pixels on object boundaries. It's akin to "shrinking" the objects, removing small objects or irregularities from the image. After erosion, dilation is applied, compensating for the erosion step, restoring the size of largest object as background without bringing back the smaller details and noises that were removed. The last stage is hole removing, aiming at filling in gaps or holes within foreground in order to get masks, which has only two connected components. The process is shown in figure 3.5

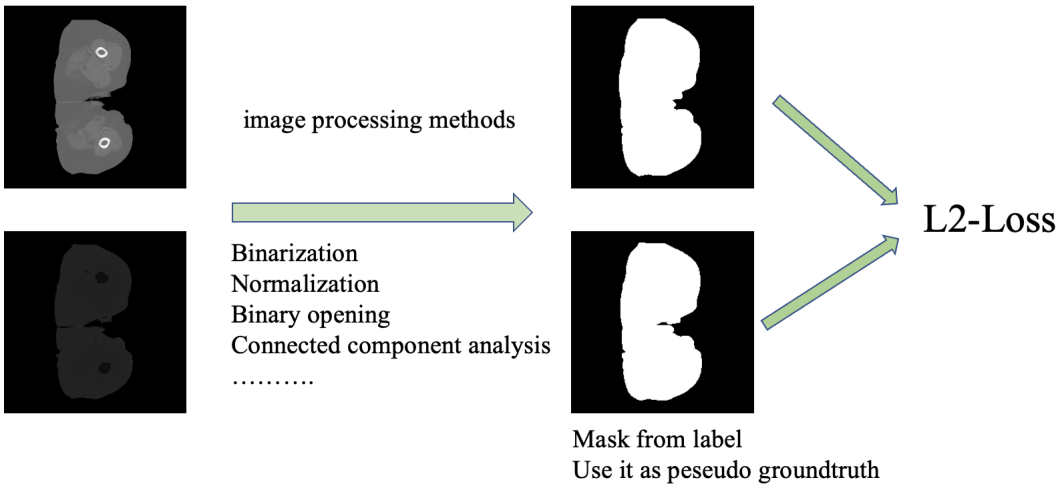

Figure 3.5.: Outputs of the U-Net

We define the mask extractor as a function $\mathcal{E} : (\mathbf{x}) \rightarrow \hat{\mathbf{m}}$. The mask loss $\mathcal{L}_{\text{mask}}$ is formulated to measure the similarity between the mask of the generated MR image $\mathcal{E}(\mathcal{G}(m))$ and the mask of the corresponding CT label $\mathcal{E}(m)$. The mask consistency loss function is then defined as:

$$\mathcal{L}_{\text{mask}} = \|\mathcal{E}(m) - \mathcal{E}(\mathcal{G}(m))\|_2^2 \tag{3.7}$$

where N represents the total number of pixels in the image, and the absolute difference between the masks is computed pixel-wise, During the training phase, we assign a weighting factor of $\lambda_{\text{mask}} = 1$ for the segmentation loss.

3.2. Pre-processing Methods

All the datasets contain 3D volumetric images. Since we perform 2D image synthesis, we split them into 2D slices from axial planes. After pre-processing these 2D image slices are fed to the network as inputs. The overview of pre-processing is shown in figure 3.6.

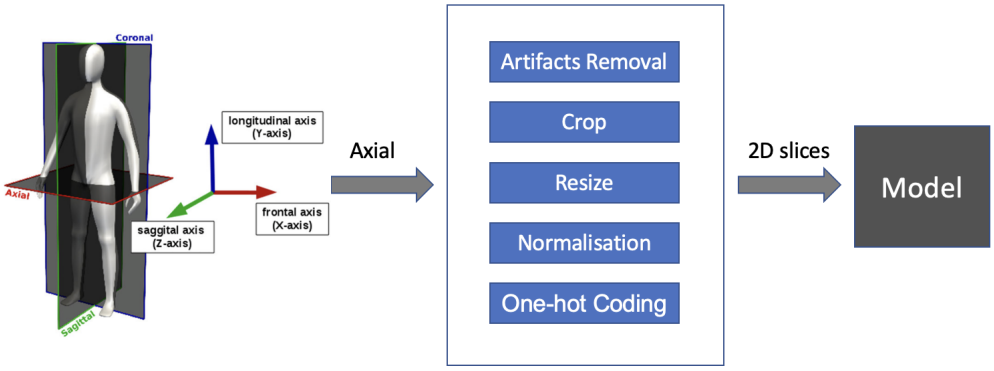

Figure 3.6.: Overview of pre-processing

3.2.1. Artifacts Removal

The process of removing artifacts from images, particularly those originating from imaging devices, is crucial for ensuring the clarity and accuracy of the main content of the image. Artifacts, which can be caused by a variety of factors including device malfunctions, noise, or software errors, often obscure or distort the true content of the image.

In our methodology, we focused on extracting and isolating significant regions from CT images. This was achieved through a series of image processing steps, primarily involving binary thresholding and connected component analysis, coupled with area-based filtering. The initial phase involved converting the grayscale CT images into binary images. This was executed by applying a thresholding operation, where pixels with intensity values greater than 45 were set to 1, and the remaining to 0, thereby distinguishing regions of interest from the background.

Subsequently, we employed the connected components algorithm with a 4-connectivity setting, and facilitated the identification and labeling of distinct connected regions within the binary image. To refine the identification of significant regions, an area-based filtering approach was adopted. Specifically, connected components with an area less than a predefined threshold of 1000 pixels were considered insignificant and hence excluded from further analysis. The remaining components, deemed significant, were then used to create a mask image. In this mask, pixels belonging to foreground region are assigned a value of 255 (white), while all other pixels were set to 0 (black).

Finally, to highlight the significant regions within the original CT images, a pixel-wise multiplication was performed between the original CT images and the generated mask. This step resulted in an image where only the significant regions are visible, effectively enhancing the features of interest while suppressing the irrelevant background.

Through this method, we are able to efficiently isolate and emphasize key structures in CT images, paving the way for more focused and detailed analysis in subsequent stages of our study.

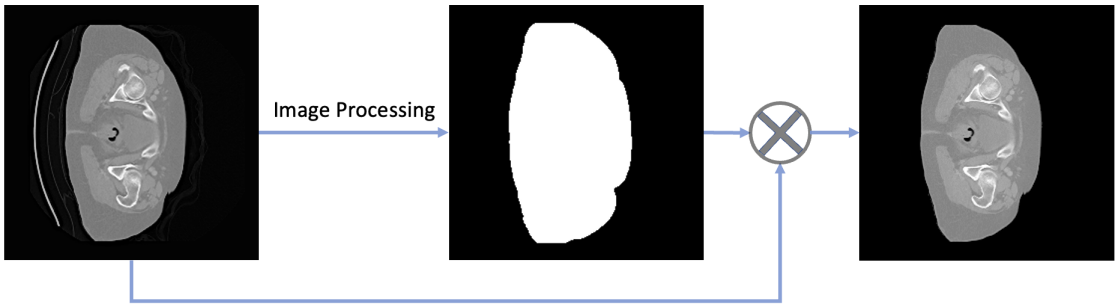

Figure 3.7.: Visualization of the process of removing background artifact

3.2.2. Data Augmentation

In medical image generation, data augmentation plays an important role in improving the robustness and generalizability of machine learning models. We implement a simple yet effective form of data augmentation: random horizontal flipping. With a 50 % chance, both the medical image and its corresponding label map are flipped horizontally. This technique helps the model to become invariant to the directionality of features, which is particularly important in medical imaging where anatomical structures can have symmetrical properties. By effectively doubling the dataset with mirrored versions of each image, the model is trained to recognize patterns and structures regardless of their orientation.

3.2.3. Normalisation, Resizing and One-hot Coding

Normalisation

We utilize Min-max normalisation methods to normalize the input images. This method scales the range of pixel value between the desired minimum and maximum values, in our case 0 and 1. One of the primary advantages of min-max normalization is that it preserves the relationships among the original data values while normalizing the range. It scales the data within a specified range, enhancing the stability and performance of the learning algorithm, preserves the original data distribution, which is crucial for algorithms that assume uniform distribution. It's beneficial for algorithms that do not assume any distribution of data, especially for neural networks. It's also simple to implement and understand. The formula for min-max normalization for a value x in a dataset is given by:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \cdot (new\_max - new\_min) + new\_min \tag{3.8}$$

x_{norm} is the normalized value. x is the original value. x_{min} and x_{max} are the minimum and maximum values of the feature, respectively. new_min and new_max are the minimum and maximum values for the normalized data. In our case, these are set to 0 and 1, respectively.

Resizing

In the pre-processing of images for machine learning models, the resizing process is important for maintaining data integrity and optimizing model efficacy. Preserving the origi-

nal aspect ratio of the images is of great importance to avoid any distortion; therefore, for all datasets, we maintained a consistent aspect ratio of 1.0. Specifically, for the AutoPET dataset, images were cropped to a size of 256×256 without any resizing, and the cropping center was adaptively selected to ensure that the region of interest remains central. For the other two datasets, a similar cropping strategy was employed before resizing the images to 256×256. To preserve image quality and fidelity, all images were stored in PNG format. By adhering to these pre-processing steps, we ensure that the quality of the images is retained, providing a solid foundation for the model's subsequent analytical or predictive operations.

One-hot Coding

The semantic maps are one-hot coded before they are entered into the model and concatenate with 3D noise. Each category is represented by a separate channel in the semantic map. In this representation, each pixel in the channel corresponds to a specific anatomical structure or region of interest in the CT image, with the value 1 indicating the presence of the category at that pixel, and 0 indicating absence. This binary encoding ensures that each channel is mutually exclusive, representing only one specific category, thus providing a clear and distinct classification of the different anatomical structures in the CT image. The one-hot encoded semantic map thus becomes a multi-channel binary image, where each channel serves as a mask for a particular anatomical structure, facilitating the precise localization and identification of different regions by the neural network during the training and inference processes. Let C be the total number of unique anatomical classes in the medical CT images. For a given pixel position (x, y) in the semantic map m, the one-hot encoded representation E is a tensor of dimensions $(C, height, width)$, where each channel i in E corresponds to class i in the semantic map. The one-hot encoding operation can be defined as:

$$E_{i,x,y} = \begin{cases} 1, & \text{if } m_{x,y} = i \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

$E_{i,x,y}$ denotes the pixel value at position (x, y) in the i-th channel of the one-hot encoded tensor E, and $S_{x,y}$ represents the class label of the pixel at position (x, y) in the semantic map S. This binary, channel-wise representation effectively encapsulates the semantic information of the CT images, rendering it an appropriate and efficient input for neural network-based analysis.

4. Experiments and Results

4.1. Datasets

We conduct our methods on three benchmark datasets - AutoPET [30], SynthRAD2023 [1] and Nakko dataset. First, we train our model in a superviesed way with CT semantic maps and CT image slices from AutoPET [30] dataset, and then we trained our Med-USIS model using SynthRAD2023 [1] dataset, which contain CT-MR pairs.

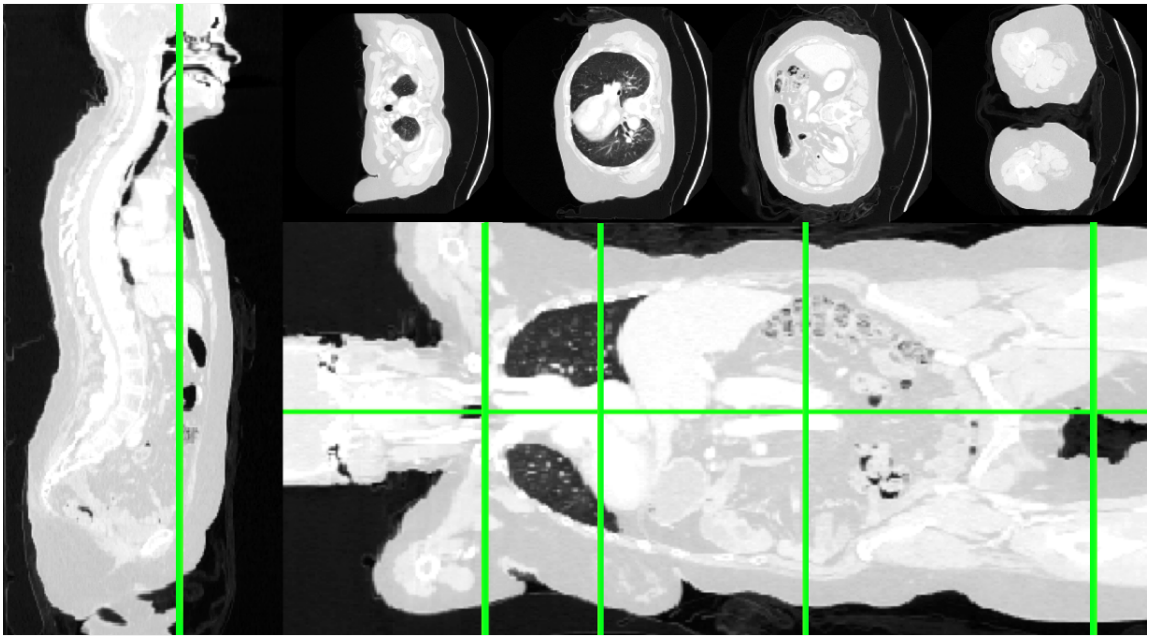

Figure 4.1.: Example of CT image from AutoPET

4.1.1. AutoPET

AutoPET dataset is a comprehensive collection of 1,014 studies (900 patients) with malignant lymphoma, melanoma, and NSCLC, along with 513 control sets without PET-positive malignant lesions, gathered at the University Hospital Tübingen between 2014 and 2018. Acquired on a Siemens Biograph mCT scanner with detailed protocol specifications, each dataset has been meticulously analyzed and manually segmented by experts. Accompanied by anonymized DICOM files, segmentation masks, primary diagnosis, and demographic data, this dataset, exemplified in the AutoPET MICCAI 2022 challenge, serves as a valuable resource for developing and training machine learning models in PET/CT image analysis.

900 patients are included in training set and 150 are included in test set. An example of CT image is shown in 4.1.

This dataset contains 37 annotated classes, which is highly imbalanced as shown in A.1. About 81 percent pixels belong to background. Among foreground classes, the top 4 classes takes more than 80 percent pixels. It is a challenging dataset for medical image synthesis and segmentation. We implement our model with paired CT images from this dataset. We randomly choose CT images from 5 percent patients as test set, CT images from 5 percent patients as validation set and the rest of images as training set.

4.1.2. SynthRAD2023

SynthRAD2023 dataset, encompassing CT, CBCT, and T1-weighted MRI images of 540 brain and pelvic radiotherapy patients from three Dutch medical centers, offers a diverse range of imaging data with subjects aged 3 to 93. Aimed at advancing sCT generation for radiotherapy planning, the collection supports the evaluation and development of image synthesis algorithms, addressing the increasing importance of medical imaging in oncology, especially in radiotherapy contexts. With comprehensive data available on Zenodo under the SynthRAD2023 collection, and presented in nifti format, this dataset stands as a pivotal resource for enhancing diagnosis, treatment planning, monitoring, and surgical planning in radiation therapy. In this dataset, the semantic maps are not included, we implement TotalSegmentator [6] to generate semantic maps. TotalSegmentator is a powerful tool, which can Robust segment 117 Anatomical Structures in CT images as shown in 4.2. We selected and merged some classes and ended up with a total of 31 classes, which are shown in A.3

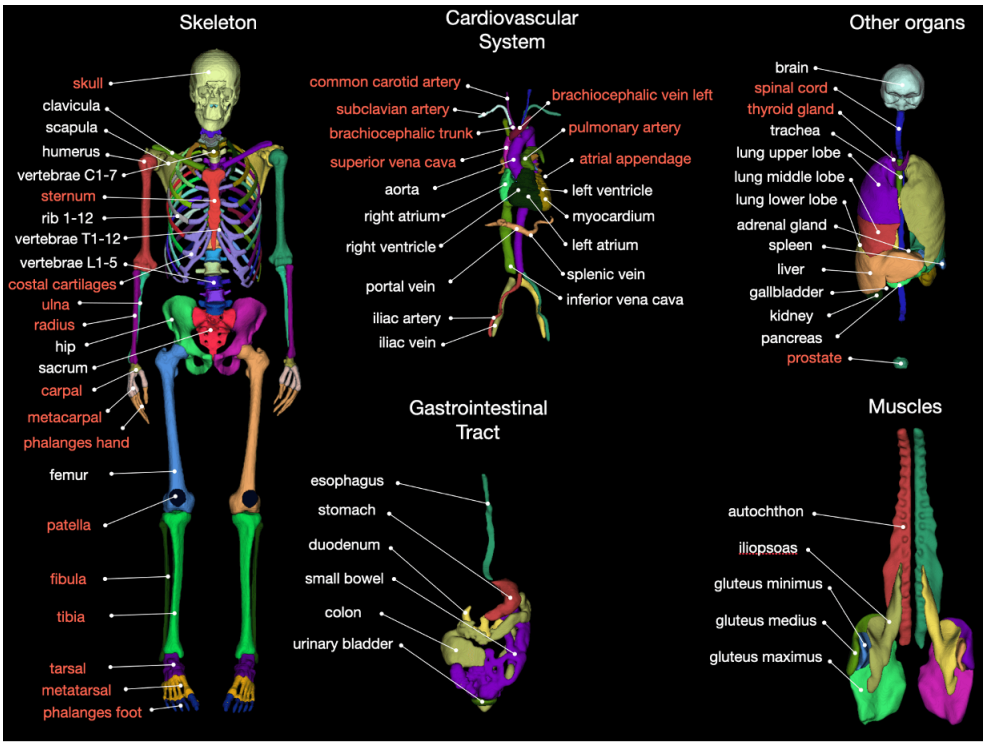

Figure 4.2.: Overview of 117 anatomical structures segmented by the TotalSegmentator [6]

4.2. Evaluation metrics

To conduct both quantitative qualitative assessment of images, various metrics have been developed to address different aspects of image quality, including error measurement, perceptual similarity, and structural integrity. Among these, Fréchet Inception Distance (FID), Root Mean Square Error (RMSE), Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity Index (SSIM), Mean Intersection over Union (MIoU)Mean Absolute Error (MAE) and Peak Signal-to-Noise Ratio (PSNR) are widely used. We use this metrics to compare the ground-truth and synthesized MR images for the qualitative comparisons. For the qualitative comparisons. The primary focus of this paper revolves around theoretical advancements. Future work will extend the evaluation to human aspects, including inviting a medical professional to evaluate. For evaluation, we set batch size to 20, which means one time we generate 20 images and measure metrics between them and take the average, the final scores will be the average of all batches.

4.2.1. Fréchet Inception Distance (FID)

The Fréchet Inception Distance (FID) is a performance metric for evaluating the quality of images generated by Generative Adversarial Networks (GANs) and other generative models. FID measures the similarity between the distribution of generated images and the distribution of real images, capturing essential aspects such as fidelity and diversity of the generated images.

FID is computed by embedding a set of generated images and a set of real images into a feature space given by a specific layer of the Inception network. Then, it calculates the Fréchet distance between the two resulting Gaussian distributions. Mathematically, the FID between the real image distribution X with mean μ_x and covariance Σ_x, and the generated image distribution $\hat{X}$ with mean $\mu_{\hat{x}}$ and covariance $\Sigma_{\hat{x}}$ is defined as:

$$\mathrm{FID}(\mathrm{X}, \hat{X}) = \|\mu_x - \mu_{\hat{x}}\|_2^2 + \mathrm{tr}(\Sigma_x + \Sigma_{\hat{x}} - 2(\Sigma_x\Sigma_{\hat{x}})^{\frac{1}{2}}) \tag{4.1}$$

where $\|\cdot\|_2$ denotes the L2 norm, and tr denotes the trace of a matrix. A lower FID indicates better quality of generated images, suggesting that the generated images are more similar to the real images in terms of both content and diversity.

4.2.2. Structural Similarity Index (SSIM)

SSIM is a perceptual metric that quantifies the visual impact of three characteristics of an image: luminance, contrast, and structure. Unlike RMSE that treats all pixel discrepancies equally, SSIM considers changes in structural information, luminance, and contrast. The SSIM index for two windows x and $\hat{x}$ of common size $N \times N$ is defined as:

$$\mathrm{SSIM}(\mathrm{x}, \hat{x}) = \frac{(2\mu_x\mu_{\hat{x}} + c_1)(2\sigma_{x\hat{x}} + c_2)}{(\mu_x^2 + \mu_{\hat{x}}^2 + c_1)(\sigma_x^2 + \sigma_{\hat{x}}^2 + c_2)} \tag{4.2}$$

where μ_x, $\mu_{\hat{x}}$ are the average of x and $\hat{x}$; σ_x^2, $\sigma_{\hat{x}}^2$ are the variance of x and $\hat{x}$; $\sigma_{x\hat{x}}$ is the covariance of x and $\hat{x}$; and c_1, c_2 are constants to stabilize the division with weak denominator. SSIM value ranges from -1 to 1, where 1 indicates perfect similarity.

4.2.3. Learned Perceptual Image Patch Similarity (LPIPS)

LPIPS [31] is a recent metric that leverages deep learning to assess perceptual similarity between images. Unlike traditional metrics that focus on pixel-level differences, LPIPS uses a neural network to compare image patches, capturing more complex characteristics and structures. It better correlates with human judgment of visual similarity. The metric involves a pre-trained deep network that extracts features from image patches and compares them. In our cae, We use a VggNet [32] as our feature extractor, which is trained on ImageNet [33] dataset. The general notion of LPIPS can be expressed as:

$$\text{LPIPS} = d(\phi(x), \phi(\hat{x}_i)) \tag{4.3}$$

where ϕ represents the feature extractor (usually layers of a deep neural network), and d is a distance metric (e.g., Euclidean distance) between the feature representations of the ground truth image x_i and the generated image $\hat{x}_i$. The detailed implementation might vary, and specific layers and weights might be chosen based on the network architecture used.

4.2.4. Mean Intersection over Union (MIoU)

In our evaluation framework, the Mean Intersection over Union (MIoU) metric is employed as a pivotal measure to assess the segmentation performance of the model. MIoU provides an aggregate measure of overlap between the predicted segmentation and the ground truth across all classes, making it particularly essential in multi-class segmentation tasks to ensure a balanced assessment across various categories. Mathematically, MIoU is formulated as follows:

$$\text{MIoU} = \frac{1}{N} \sum_{i=1}^{N} \text{IoU}_i \tag{4.4}$$

where N is the number of classes, and IoU_i represents the Intersection over Union for the i^{th} class, computed as:

$$\text{IoU}_i = \frac{\text{Area of Overlap}_i}{\text{Area of Union}_i} \tag{4.5}$$

In this formulation, Area of Overlap_i denotes the intersection area between the predicted segmentation and the ground truth for class i, and Area of Union_i represents the union of the predicted and actual segmented areas for the same class. In our research, we concentrate on evaluating the quality of the generated CT images because we don't have a specialized network for segmenting MRI images. For this purpose, we used a 2D CT image segmentation network that we built from the ground up, relying on the nnUNet framework. This method

allows us to precisely assess how well our model performs in segmenting CT images, and it also points out the importance of developing a similar, effective segmentation tool for MRI images in future studies.

4.2.5. Mean Absolute Error (MAE)

To assess the quality of generated images in our study, we also employ the Mean Absolute Error (MAE) as a key metric. MAE is a standard measure used in statistical analysis to quantify the difference between two continuous variables. In the context of image generation, it calculates the average magnitude of errors between the generated image and the ground truth, pixel by pixel. The MAE is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i| \tag{4.6}$$

where $\hat{x}_i$ represents the pixel values of the generated image, x_i denotes the pixel values of the ground truth image, and n is the total number of pixels in the images. A lower MAE score indicates higher fidelity of the generated image to the ground truth, reflecting better image synthesis quality.

4.2.6. Peak Signal-to-Noise Ratio (PSNR)

PSNR is commonly used to measure the quality of reconstruction of lossy compression codecs and image synthesis models. It compares the maximum possible power of a signal (original image) to the power of corrupting noise (error in the processed image). The formula for PSNR is:

$$\text{PSNR} = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \tag{4.7}$$

where MAX_I is the maximum possible pixel value of the image, and MSE is the mean squared error between the original and the processed image. Higher PSNR generally indicates better image quality, correlating with a lower error.

Each of these metrics provides insights into different aspects of image quality, facilitating comprehensive analysis and evaluation in various image processing tasks.

4.2.7. Root Mean Square Error (RMSE)

RMSE is a standard way to measure the error of a model in predicting quantitative data. In the context of image processing, it quantifies the pixel-wise difference between the original and the processed image. It is defined as the square root of the average of the squares of the differences between corresponding pixel values of the two images. The formula for RMSE is:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i - x_i)^2} \tag{4.8}$$

4.3. Experimental Setup

In our approach, we drew inspiration from the OASIS framework, specifically applying an exponential moving average to the weights of the generator, with a decay factor set at 0.9999. We standardized the resolution of images to 256×256 across all three datasets. The training was executed with a batch size of 4 on a single RTX A5000 GPU with 24 G memory. For the optimization process in all our experiments, we employed the ADAM algorithm, with the momentum terms β_1 and β_2 set to 0 and 0.999, respectively, and a learning rate of 0.0001. We also set the coefficients λ_{seg} and λ_{mask} to 1.0, aiming for a balanced impact on the overall performance of our model.

4.4. Ablation Study

In table 4.1. We trained the model on AutoPET dataset in both paired and unpaired paradigm. This experiments aim at transferring CT labels to CT images, since MR images are not included in AutoPET dataset. We perform the ablation study to analyze the effect of different components of our models. All the experiments have the same images resolution of 256×256 for both input and generated images. To assess the quality of the CT images produced by our model, we trained a nnUNet segmentation network from scratch with CT slices from AutoPET. This network was specifically used to calculate the Mean Intersection over Union (MIoU). The results were promising, with the nnUNet achieving a Dice score of 0.78 on the validation set. The training process is shown in A.6. Due to the class imbalance present in the AutoPET dataset, the IoU (Intersection over Union) metric exhibits variance across different classes. Specifically, the IoU values for minor classes tend to be lower, reflecting the challenge in accurately segmenting less represented categories. However, for major, more prevalent classes, the IoU values are notably higher, often exceeding 0.75 as shown in A.2.

Table 4.1.: Ablation study on AutoPET

Exp	Experiment Details				Results				
	Unpaired	Generator OASIS	Generator Wavelet	$\mathcal{L}_{mask}$	FID	LPIPS	SSIM	RMSE	PSNR
Exp-1		✓			15.83	0.27	0.9714	0.923	15.39
Exp-2		✓		✓	**5.67**	0.22	0.9713	0.45	19.83
Exp-3			✓		7.29	0.06	**0.9995**	**0.06**	**24.89**
Exp-4			✓	✓	10.68	**0.05**	0.9995	0.06	23.27
Exp-5	✓		✓		10.76	0.26	0.9283	0.21	13.82

We have included Exp 1-4 In table 4.1 in the comparison as an upper bound to our unpaired model, in order to identify improvement opportunities.

Table 4.2.: Ablation study on SynthRAD2023

Exp	Experiment Details				Results				
	Unpaired	Generator OASIS	Generator Wavelet	$\mathcal{L}_{mask}$	FID	LPIPS	SSIM	RMSE	PSNR
Exp-1	✓	✓			60.93	0.15	0.9983	0.12	18.54
Exp-2	✓	✓		✓	59.97	0.16	0.9980	0.12	18.09
Exp-3	✓		✓		**51.92**	0.15	0.9983	0.12	18.69
Exp-4	✓		✓	✓	54.53	**0.15**	0.9984	0.12	**18.77**
Exp-5		✓		✓	60.70	0.17	0.9978	0.13	17.69
Exp-6			✓	✓	68.35	0.17	**0.9985**	**0.11**	**19.08**

In table 4.2. We trained the model on SynthRAD2023 dataset in both paired and unpaired paradigm. This experiments aim at transferring CT labels to MR images. We perform the ablation study to analyze the effect of different components of our models. All the experiments have the same images resolution of 256×256 for both input and generated images. We first train our model using unpaired data (Exp 1-5 In table 4.2). Then train the model using paired data. But in our case, the corresponding CT and MR still have differences especially on the boundary as shown in figure 4.3, so the results are not as good as the Exp 1-4 in table 4.1 where the semantic maps and images are exactly corresponding to each other. They even have a higher FID than the unpaired paradigm.

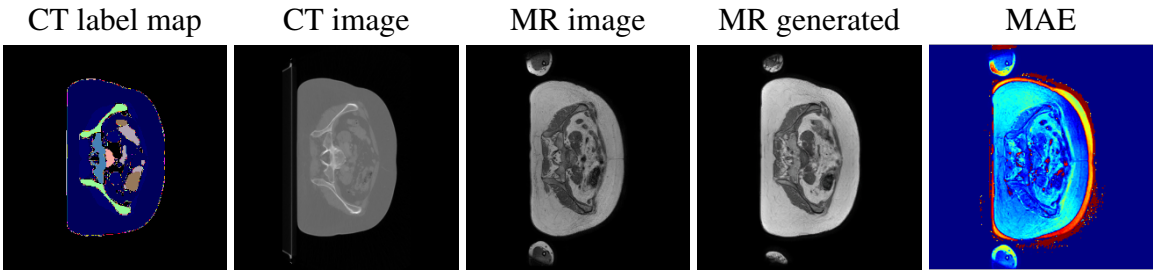

Figure 4.3.: MR generation results and MAE

4.4.1. Ablation Study on the Generator Type

In our first experiment, we utilized the OASIS generator as the foundational architecture for image generation, which shows an FID of 15.83. Observations from the outcomes, as shown in figure A.1, indicate a significant variation in the shapes of the generated images contingent on the input noise. Notably, the generator appears to struggle with aligning the shapes in the generated images shape with the expected labels, resulting in a lack of consistent structural fidelity. This variability points to a potential challenge in the model's capacity to capture and replicate the precise structural essence inherent in the target domain. To address the observed inconsistency in the structural alignment of the generated images, we incorporated a mask loss into our model. This loss function is predicated on the assertion that the foreground of the generated image should be congruent with that of the semantic map. The introduction of mask loss aims to penalize the discrepancies between the generated image and the semantic

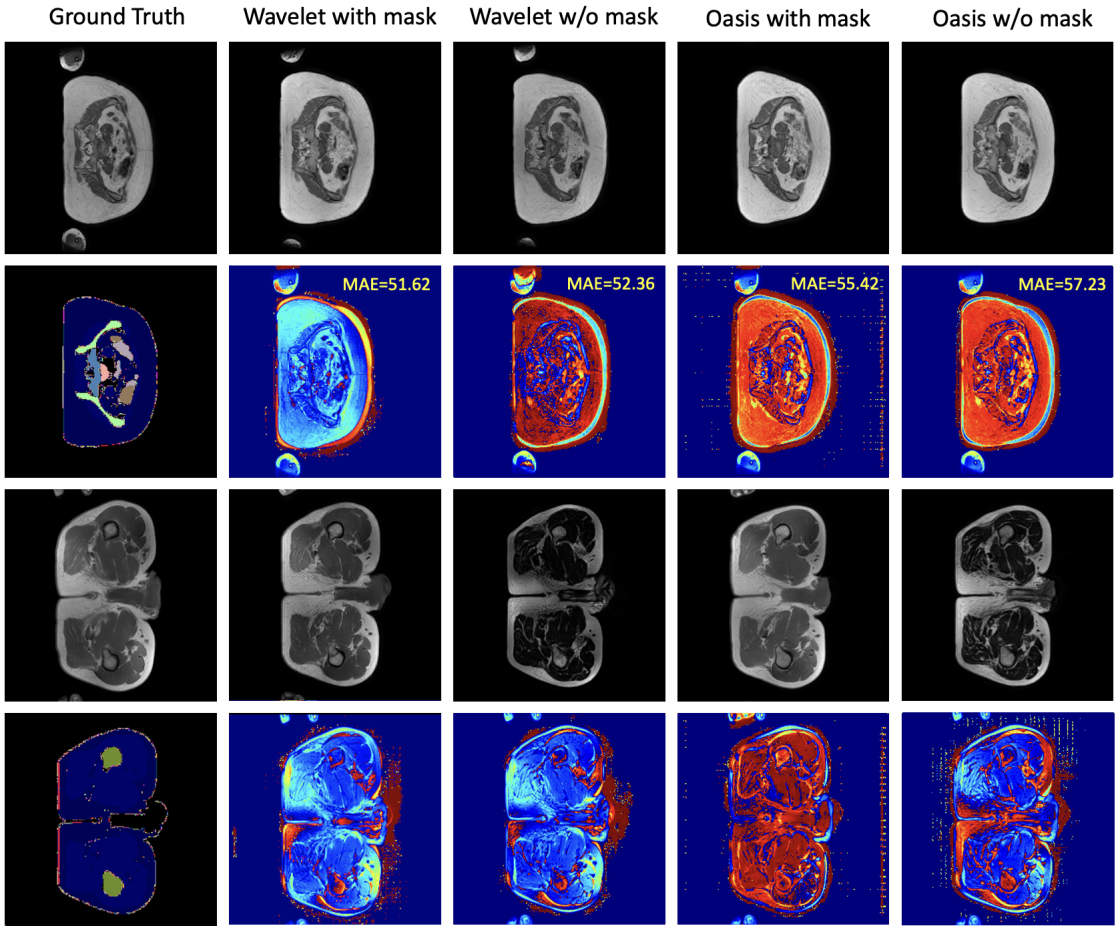

Figure 4.4.: Comparison of our methods using MAE

map, thereby encouraging the model to preserve the structural integrity of the target objects. The integration of mask loss into our model yielded notable improvements in the subsequent experiment. The generated images exhibited enhanced consistency in shape, more closely mirroring the structure delineated in the semantic maps, as shown in figure A.2.

However, the wavelet generator demonstrated a superior capability in capturing and aligning the shapes within the generated images. This observation is substantiated not only by the qualitative assessment of the generated images but also through the quantitative results. The wavelet-based method offers a more refined and nuanced approach to maintaining structural integrity, effectively capturing the intricate details and contours of the target shapes. The integration of mask loss, while beneficial in traditional generative settings, appears to have a limited impact when coupled with the wavelet-based model. Thereby making it a less crucial component in the Wavelet generator framework. Its ability to operate at multiple resolutions allows it to preserve both the global structure and the fine details, resulting in generated images that exhibit a higher degree of shape alignment and visual coherence. The improved performance, as evidenced by the enhanced FID scores and the visual quality of the generated images, underscores the potential of wavelet transformation as a pivotal component in our generative modeling framework.

Figure 4.6 presents a qualitative analysis comparing the outputs from our two generators. It

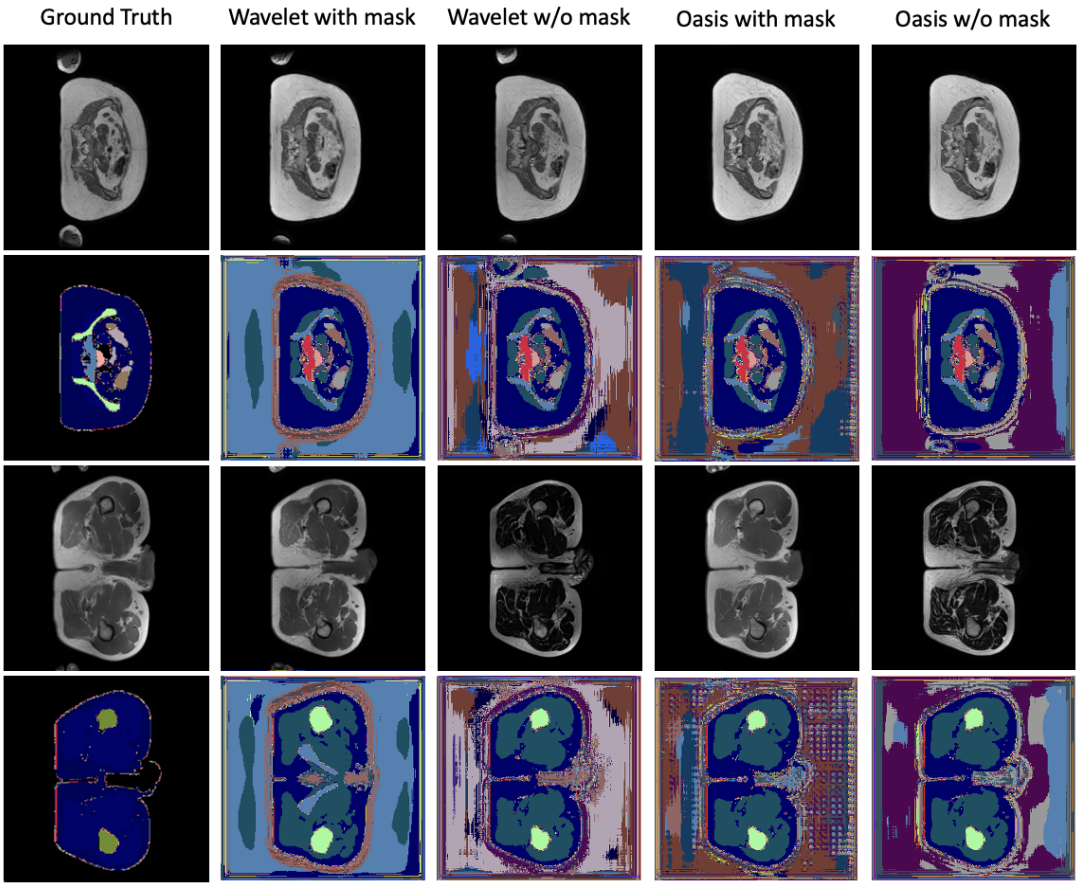

Figure 4.5.: Output of the U-Net

is evident that the images produced by the Wavelet generator possess superior quality, with more refined textures and edges. This demonstrates that the Wavelet generator is better suited for the task of medical image synthesis. Figure 4.5 displays the segmentation results from our U-Net based model. The outputs reveal that while the images generated by the OASIS generator lead to rather noisy label maps, those produced by the Wavelet generator yield more precise and clear segmentation outcomes. Figure 4.4 illustrates the MAE computed between the ground truth images and those generated by our models. MAE, a measure of average pixel-wise difference, offers insights into the fidelity of generated images compared to the original. Despite some noticeable misalignments along the boundaries in all cases, the images produced by the Wavelet generator exhibit a significantly lower MAE compared to those from the OASIS generator. This observation holds true irrespective of the presence or absence of mask loss in the model configurations. The lower MAE values associated with the Wavelet generator prove its superior performance in replicating finer details and maintaining closer alignment with the ground truth, highlighting its effectiveness for our image synthesis tasks.

4.4.2. Ablation Study on the Mask Consistency Self-Supervision

To address the observed inconsistency in the structural alignment of the generated images, we incorporated a mask loss into our model. This loss function is based on the assertion that

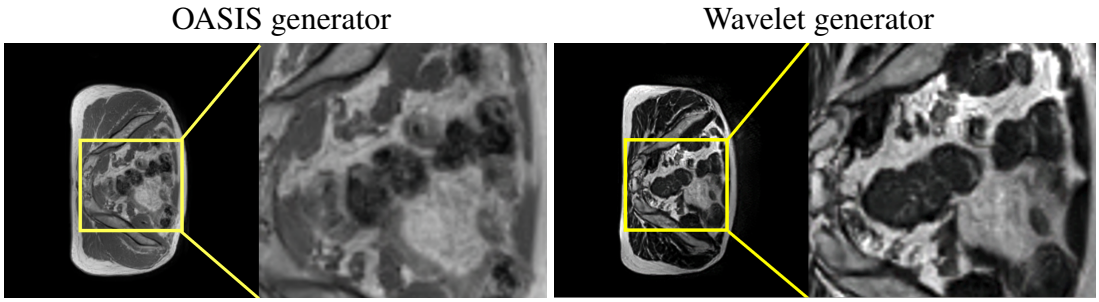

Figure 4.6.: Qualitative comparison of our two generators.

the foreground of the generated image should be congruent with that of the semantic map. The introduction of mask loss aims to penalize the discrepancies between the generated image and the semantic map, thereby encouraging the model to preserve the structural integrity of the target objects. The integration of mask loss into our training regimen yielded notable improvements in the subsequent experiment, especially in CT label to CT task with Oasis generator. The results are shown in figure A.1 and A.2. The generated images exhibited enhanced consistency in shape, more closely mirroring the structure delineated in the semantic maps. Moreover, this adjustment manifested in the quantitative metrics, with FID witnessing a substantial reduction to 5.67 as shown in table 4.1 Exp-1 and Exp-2. This improvement in FID proves the efficacy of mask loss in enhancing the shape consistency of the generated images, aligning them more accurately with the prescribed semantic contours.

However, This significant improvement does not appear in Wavelet generator according to Exp-3 and Exp-4 in table 4.2. The incorporation of the wavelet transformation in our approach has demonstrated a superior capability in capturing and aligning the shapes within the generated images. This observation is substantiated not only by the qualitative assessment of the generated images but also through the quantitative results. The wavelet-based method offers a more refined and nuanced approach to maintaining structural integrity, effectively capturing the intricate details and contours of the target shapes. Its ability to operate at multiple resolutions allows it to preserve both the global structure and the fine details, resulting in generated images that exhibit a higher degree of shape alignment and visual coherence. The improved performance, as evidenced by the enhanced FID scores and the visual quality of the generated images, underscores the potential of wavelet transformation as a pivotal component in our generative modeling framework.

The integration of mask loss, while beneficial in OASIS generator, appears to have a limited impact when coupled with the wavelet-based model. This observation suggests that the wavelet model, by its inherent design and operational mechanics, effectively encapsulates and retains the structural nuances and shape alignments within the generated images. The multi-resolution nature of wavelet transformation enables a profound capture of both coarse and fine details, potentially rendering the additional imposition of mask loss redundant or less impactful. The outstanding capability of the wavelet model to inherently preserve and emphasize critical structural features might diminish the relative utility or enhancement offered by mask loss, thereby making it a less crucial component in the wavelet-based generative framework.

4.4.3. Ablation Study on 3D Noise Input

In our generative model, we incorporate 3D noise as part of the input to enhance the diversity of generated images and prevent the model from producing identical outputs for similar inputs. This noise input contributes to the creation of more realistic medical images and helps the model avoid the tendency to replicate training data. By introducing randomness through noise, the model becomes more robust in handling new and previously unseen images and is capable to do multimodal generation. However, the use of noise input introduces an element of unpredictability, as illustrated in figure 4.7. This figure demonstrates significant variations in the visibility of images generated from the same label when different noise inputs are applied. These challenging scenarios, referred to as "bad cases," occur when noise inputs with extreme values, which have low probability are utilized. This can be expressed as: $z \sim P(z)$, $z \in$ tail of P. where z is the sampled noise vector and $P(z)$ is the probability distribution from which z is drawn. The exploration of low-probability samples through noise input sampling represents an excursion into the tail regions of the data distribution. This scenario occurs when the noise vector, sampled from a probability distribution like a Gaussian, falls into the areas less represented in the distribution, typically the tails.

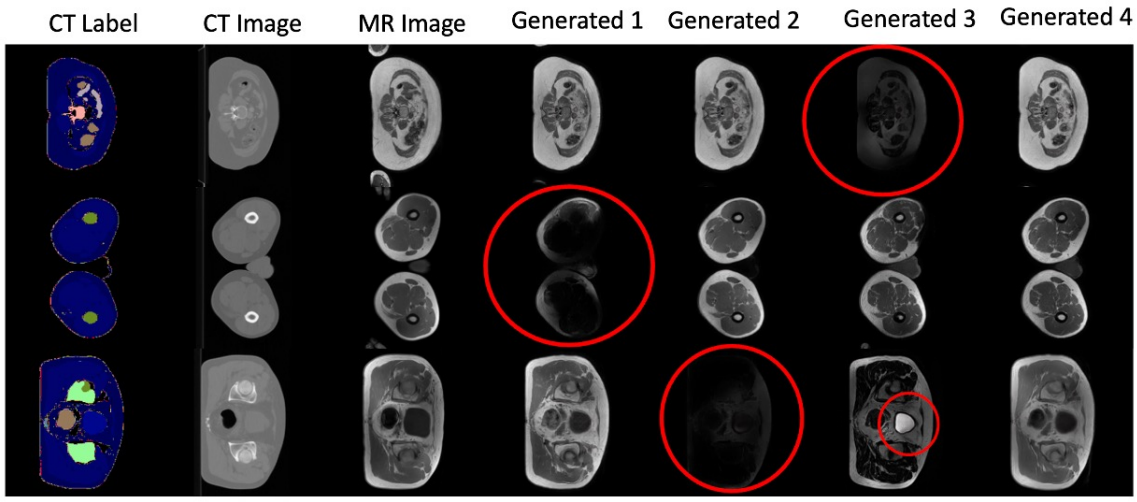

Figure 4.7.: Failure cases

The generation of such low-probability samples is significant for several reasons. Firstly, it tests the ability of the GAN to generalize beyond the common patterns found in the training data. Secondly, it is a measure of the diversity that the GAN can introduce in its outputs, which is crucial for applications requiring a wide range of variations. However, this also poses a challenge as the generated samples from these regions might be less realistic or coherent, owing to the limited training data representing such rare occurrences. Therefore, while the exploration of these low-probability regions is essential for understanding the full capacity of the GAN, it also requires careful consideration in balancing the realism and diversity of the generated samples.

In some instances, the variance can be so extreme that it leads to rare cases where the generated images almost vanish, highlighting the complexities in managing the impact of noise on the image generation process. This happens more often with OASIS generators, which means output diversity of OASIS generator is more limited than Wavelet generator

Truncated normal distribution as noise input during inference

To mitigate the occurrence of "bad cases," we employ a strategy of generating images with noise inputs sampled from a truncated normal distribution. This approach is visually depicted in figure 4.8, where a 2D representation of the truncated normal distribution is presented. Using a truncated normal distribution as noise input in generative models introduces controlled variability and enhances model stability without the risk of extreme values. This approach ensures that the noise input remains within a certain range [-a, a], thereby preventing the generation of unrealistic samples and reducing the interference of noise during the model's inference process. The use of truncated normal distribution is favored for its ease of implementation and debugging, with adjustable parameters to suit specific application scenarios. Overall, it provides a balanced method to incorporate necessary randomness while maintaining the generation process's control and stability, essential for producing high-quality, credible outputs.

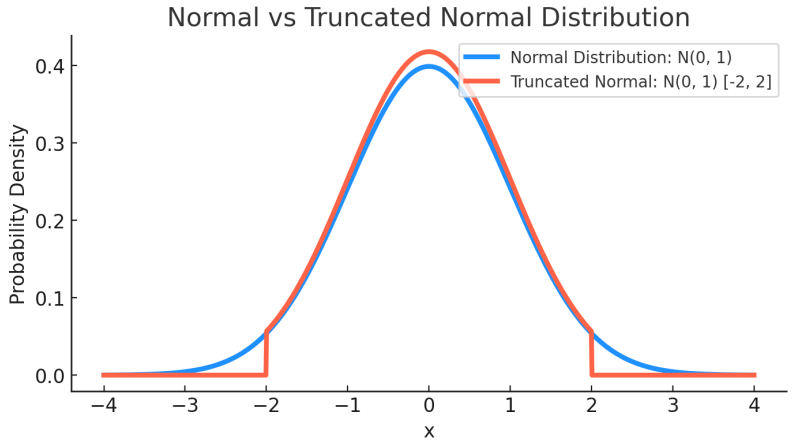

Figure 4.8.: Normal vs Truncated Gaussian distribution

we contrast the probability density functions of the standard normal distribution with those of the truncated normal distribution to examine the impact of truncation on data distribution. As illustrated in the figure 4.8, the standard normal distribution (represented by the deep blue line) exhibits the characteristic bell curve across the entire range of values, with the central point at the mean (0). In contrast, the truncated normal distribution (represented by the red line) displays a similar shape to the standard normal distribution within the specified interval [-2, 2], but values outside this interval are truncated to 0. This truncation strategy restricts the range of data values, thereby introducing randomness into generative models while controlling the occurrence of extreme values and enhancing model stability. This comparison demonstrates the potential advantages of using the truncated normal distribution as noise input in specific application scenarios, particularly when it is necessary to constrain the input distribution to avoid unrealistic outputs.

During the training process, noise inputs are sampled from a normal distribution. After training, when generating images, these noise inputs are sampled from a truncated normal distribution with varying sample interval [-a, a]. The results are presented in table 4.3.

The table 4.3 shows how varying the truncation levels (sample interval) affects different quality metrics of images generated by our Wavelet model. As sample interval increases, there's

Table 4.3.: Ablation on sample interval

Interval [-a, a] \ Metrics	FID	LPIPs	SSIM	RMSE	PSNR
[-0.001, 0.001]	62.93	0.14	0.9987	0.11	19.50
[-0.1, 0.1]	62.31	0.14	0.9987	0.11	19.49
[-0.2, 0.2]	59.27	0.14	0.9986	0.11	19.37
[-0.5, 0.5]	57.77	0.14	0.9987	0.11	19.42
[-0.8, 0.8]	54.27	0.14	0.9986	0.11	19.37
[-1, 1]	52.76	0.14	0.9986	0.11	19.34
[-1.5, 1.5]	51.12	0.14	0.9986	0.11	19.23
[-2, 2]	50.76	0.14	0.9985	0.11	19.1
[-3, 3]	51.32	0.14	0.9985	0.11	19.01
[-∞, ∞]	51.92	0.15	0.9984	0.12	18.77

a general trend of improvement in the FID as shown in figure 4.9, indicating better quality of generated images, as they become more similar to real images. A broader range of noise input may help the model generalize better, meaning it can effectively handle a wider variety of input variations, thereby producing more diverse but still high-quality images. This improvement continues up to a certain point, after which the quality slightly declines, as suggested by a small increase in FID at 'a' = 3. This decline of image quality can indeed be attributed to failure cases in the generated images, which are likely a result of the characteristics of the noise input at this higher truncation level, such as 'a' = 3, the range of noise input includes values that occur with lower probability in a standard normal distribution. This means the model is now exploring areas of the input space that it encountered less frequently during training. The generative model might not have learned effective representations for these less common, more extreme noise inputs. As a result, the images generated from such noise inputs might be less realistic or have aberrant features, leading to the observed increase in FID. We argue that an FID minima occurs as we increase the sampling interval, which can be used to improve the quality of generated images.

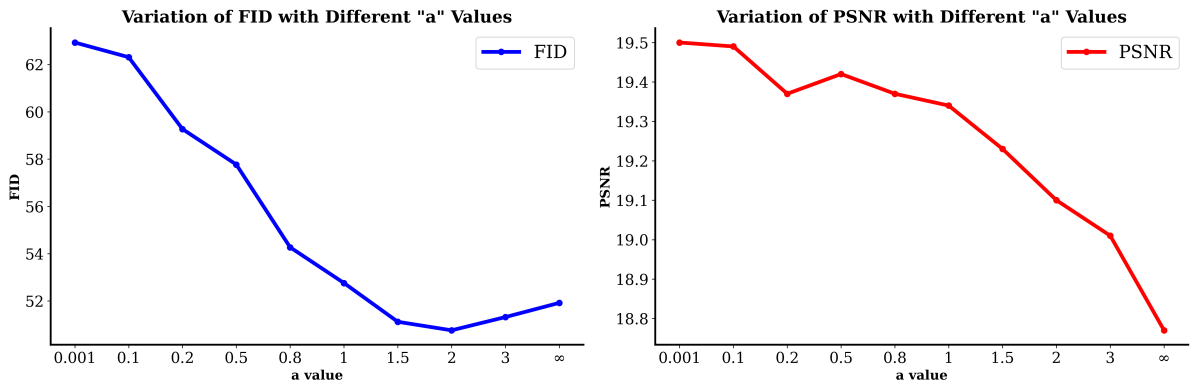

Figure 4.9.: Variation of FID and PSNR with "a" value

Other metrics such as LPIPs, SSIM, RMSE, and PSNR remain relatively stable across different truncation ranges. This indicates that while the overall similarity to real images (as

measured by FID) improves with a broader range of noise inputs, the perceptual similarity, structural integrity, average error, and signal-to-noise ratio of the images are less affected by changes in the truncation range. The slight decrease in PSNR at higher 'a' values suggests a minor degradation in image quality, but the changes are not substantial, indicating that the model maintains a consistent performance across a variety of truncation levels.

One to one mapping without 3D noise input during training

In the previous case with noise input, the generator defines a mapping function $\mathcal{G} : (\mathbf{z}, \mathbf{m}) \rightarrow \hat{\mathbf{x}}$, where $\mathbf{z}$ is the noise vector, $\mathbf{m}$ is the label, and $\hat{\mathbf{x}}$ is the set of generated images. This mapping accounts for the randomness of the noise and the specificity of the given label. We trained a model with Wavelet generator in parallel that has no noise inputs. Without noise inputs, the model's output for a given label is deterministic. In this case, the generator defines a one to one mapping function $\mathcal{G} : (\mathbf{m}) \rightarrow \hat{\mathbf{x}}$. For each label $\mathbf{m}$ there is a unique image $\hat{\mathbf{x}}$ such that $\mathcal{G}(\mathbf{m}) = \hat{\mathbf{x}}$. This is a direct mapping without the influence of noise input. This means that for each specific input label, the model will consistently produce the same output image. This can be particularly useful in scenarios where consistency and predictability of outputs are desired. This method could be particularly beneficial in medical imaging, where you might want to generate specific types of images based on clear, defined criteria, or in scenarios where reproducibility is crucial. The comparison of metrics is shown in 4.4.

Table 4.4.: Ablation on 3D noise

3D noise input	FID	LPIPs	SSIM	RMSE	PSNR
w	51.92	0.15	0.9984	0.12	18.77
w/o	60.35	0.26	0.9951	0.19	14.19

In our ablation study, summarized in Table 4.4, we investigated the impact of 3D noise input on the quality of generated images. The results clearly demonstrate the benefits of incorporating 3D noise. When 3D noise is used, the images exhibit a significantly lower FID of 51.92, compared to 60.35 when it is omitted. This indicates a closer resemblance of the generated images to real images in terms of overall features. Moreover, the inclusion of 3D noise leads to a better LPIPs score of 0.15, suggesting a reduced perceptual difference from real images. The SSIM also improves from 0.9951 to 0.9984 with 3D noise, indicating enhanced structural fidelity. Additionally, the RMSE is lower at 0.12, and the PSNR is higher at 18.77 when using 3D noise, further attesting to the improved clarity and quality of the generated images. These findings prove the efficacy of 3D noise input in enhancing the overall quality, perceptual similarity, structural integrity, and clarity of generated images.

In summary, training a model in this manner could offer advantages in terms of output consistency and control, making it suitable for applications requiring high precision and reproducibility. However, it might also limit the quality and diversity of the generated images compared to models that incorporate noise inputs.

5. Conclusion and Outlook

This section summarizes the thesis research on the Unsupervised Semantic Image Synthesis for Medical Imaging and describes the direction of future work.

5.1. Conclusion

In this study, we propose a framework Med-USIS for unsupervised semantic image synthesis for medical imaging. The architecture of the network consists of three key components: a wavelet generator, a wavelet discriminator, and a U-Net based segmentation network. Additionally, the network includes two mask extractors designed to retrieve masks from both the input label map and the generated MR image. Central to this setup is the implementation of a mask loss function, which aims to minimize the L2 distance between the two extracted masks. This approach is particularly effective in penalizing discrepancies in shape, thereby ensuring the accurate generation of images. Our aim is to generate more medical data from semantic map for the down stream tasks, such as semantic image segmentation and recognition. challenge accentuated by the complexity and sensitivity of the data involved.

The effect of our framework was validated on three challenging datasets: AutoPET and SynthRAD2023. The proposed framework achieve a promising results. According to ablation study, the Wavelet generator is at the forefront of our experiments. It incorporates wavelet transformations with the robust feature learning of deep neural networks. This design choice was motivated by the need to capture and synthesize intricate details across different scales, especially given the unbalanced nature of the medical datasets like AutoPET.

Throughout our experiments, we contended with the prevalent issue of shape inconsistency of generated images. To address this, we introduced the mask consistency approach, using basic image processing methods like Gaussian blurring and binary opening operations to refine the mask alignment between the generated MR images and CT labels. In addition, our experiments indicates the importance of balancing the output diversity and stability of generative model.

Despite these advancements, our experiments revealed that the model struggled with classifying certain classes, especially the minor classes, leading to noisy outputs of the U-Net based segmentator, especially in the background. However, the overall shapes were well-preserved, indicating the model's strength in capturing structural information, albeit with some class misalignments.

To the best of our knowledge, it is the first work to generate MR images from CT labels with unpaired training data in a unsupervised paradigm. By removing personally identifiable information and concentrating on medically pertinent, anonymized data, this method adheres to legal and ethical guidelines while also reducing the risks to privacy that come with data exposure during the processes of sharing or publishing.

5.2. Outlook

The field of Unsupervised Semantic Image Synthesis for Medical Imaging is on the verge of significant progress, with several key areas that could advance this field. One such advancement is the exploration of Transformer Architectures, particularly notable for their success in natural language processing, offer a unique advantage in medical image synthesis due to their self-attention mechanisms. These mechanisms allow the model to weigh and integrate information across the entire image, leading to a more comprehensive understanding of complex anatomical structures. This could significantly enhance the model's ability to produce detailed and accurate medical images, especially in cases where contextual understanding of the entire image is crucial.

Diffusion Models stand out for its ability to generate images through a gradual and controlled process of denoising. Starting from a random distribution, these models iteratively refine the image details. This process is particularly aligned with medical imaging, where capturing subtle gradations in tissue densities and pathological variations is essential. The potential of Diffusion Models lies in their ability to handle these nuances, offering a method to synthesize medical images that are not only realistic but also clinically relevant.

The shift to 3D image generation is another frontier. Current 2D models mainly focus on surface details, but 3D models can encapsulate the volumetric depth crucial in medical imaging. For instance, in MRI and CT scan analysis, understanding the 3D structure of an organ or a tumor is vital. Advanced 3D generative models could provide more accurate representations, aiding in diagnoses and treatment planning.

However, these advancements come with increased computational demands. Transformer Architectures and Diffusion Models, especially in 3D, require significant computational power and memory, which might limit their use in real-time applications or in settings with limited hardware capabilities. Thus, while these technologies promise to enhance medical image synthesis, their practical deployment hinges on the availability and advancement of computational resources.

A. Additionally

A.1. More generated images

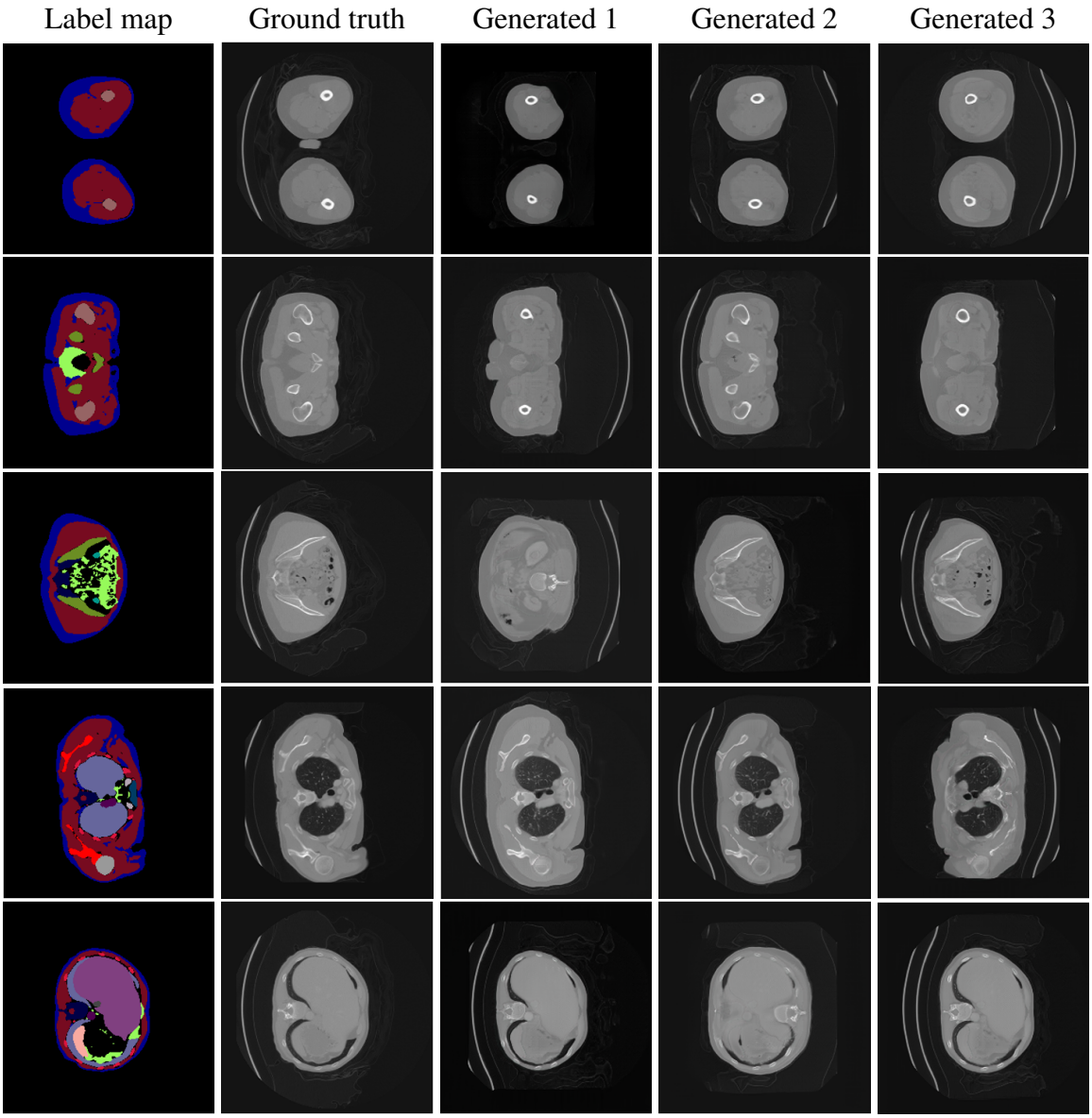

Figure A.1.: CT images from supervised model at very beginning.

Images from Exp-1 generated by OASIS [27] generator trained through pairs of CT labels and CT images from AutoPET [30] dataset. First row is label map which is colorized, second

row is Ground truth images. From third to fifth row, they are generated images with different noise samples.

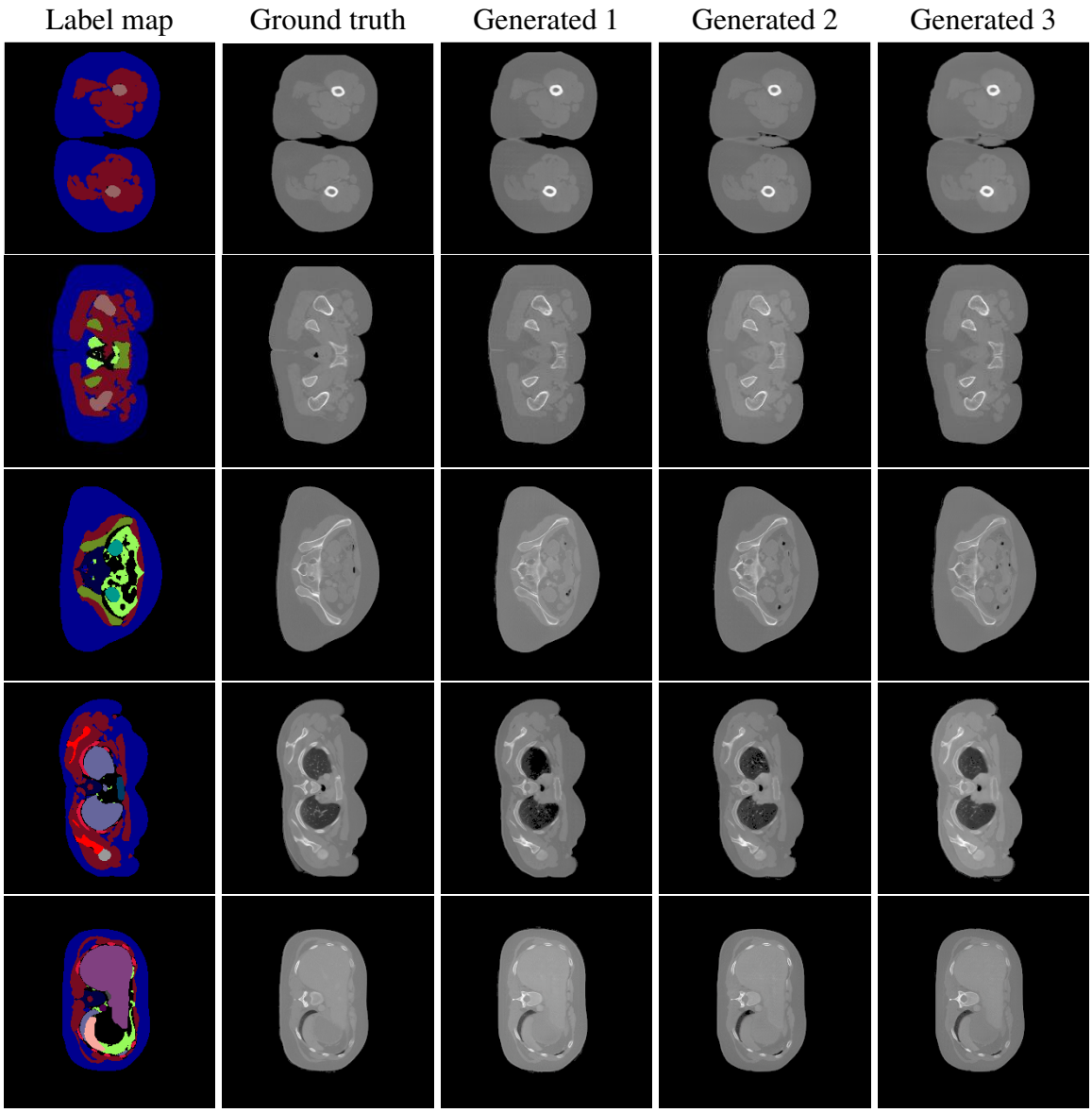

Figure A.2.: CT images from supervised model with mask loss and artifacts removement.

After removing artifacts and adding mask loss, Images from Exp-2 are generated by OASIS generator trained through CT label-CT image pairs from AutoPET [30] dataset. First row is label map which is colorized, second row is Ground truth images. From third to fifth row, they are generated images with different noise samples.

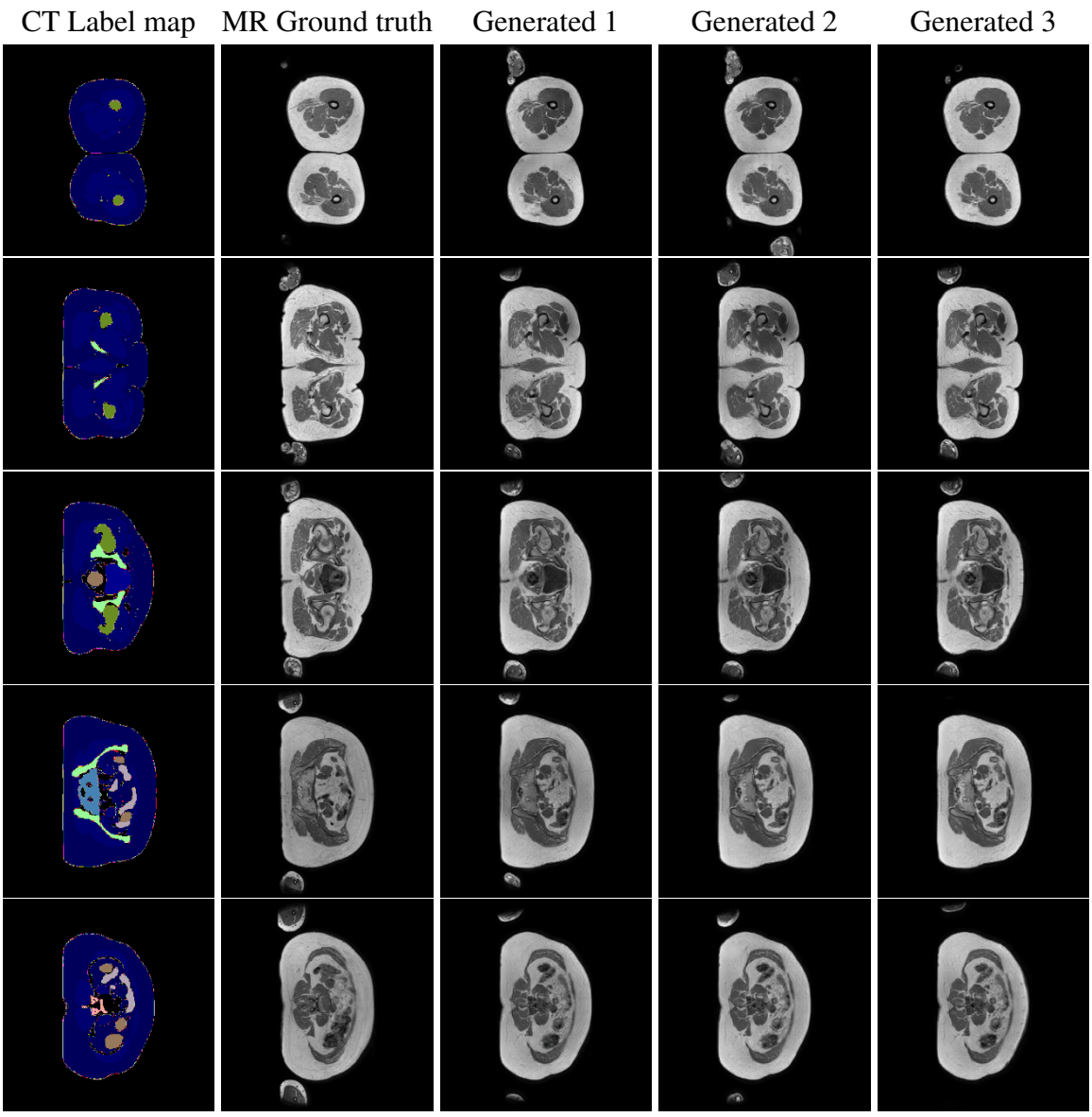

Figure A.3.: MR images from unsupervised model with Wavelet generator with mask loss.

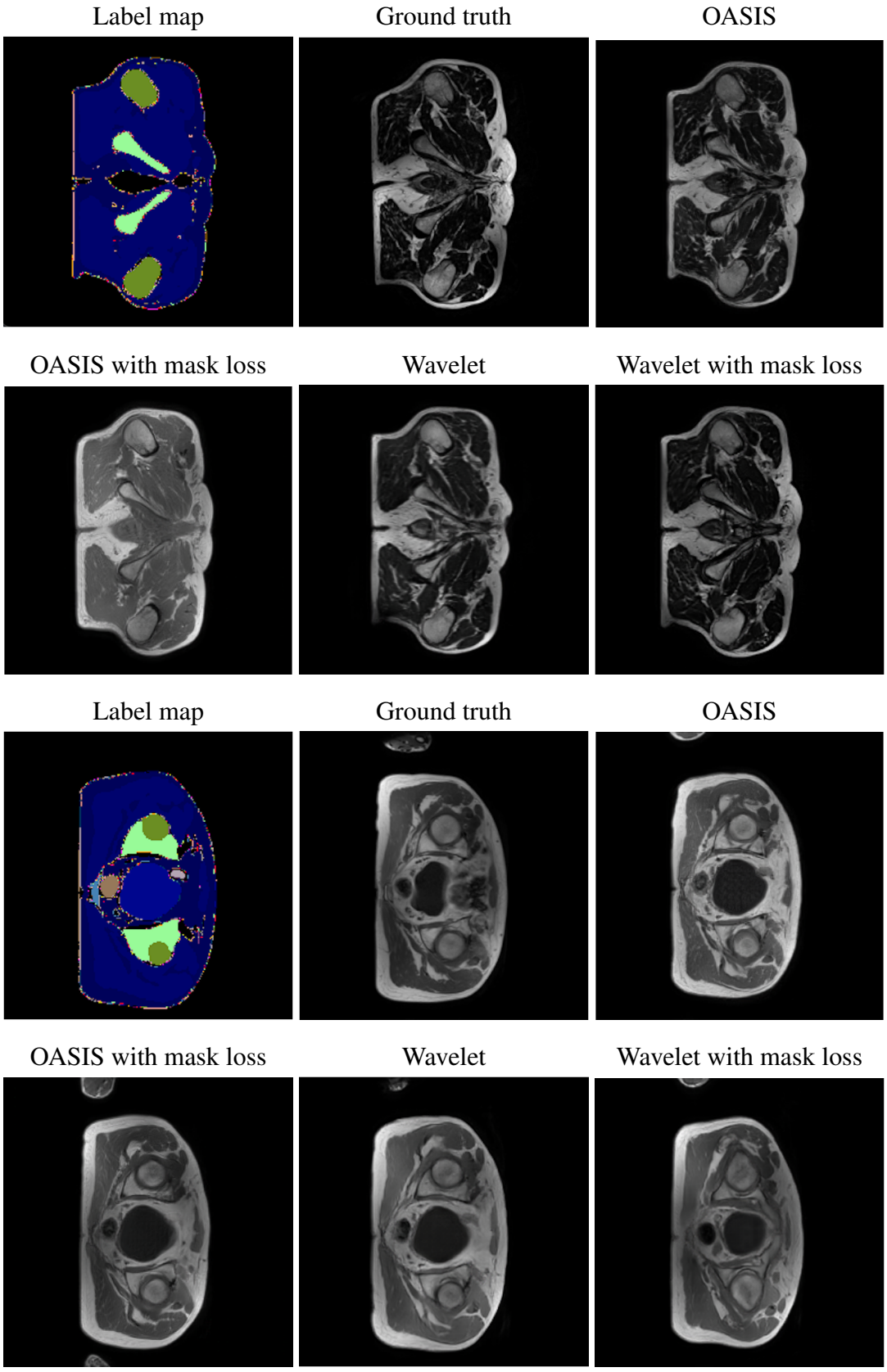

Figure A.4.: Comparison of Wavelet and OASIS generator with and without mask loss.

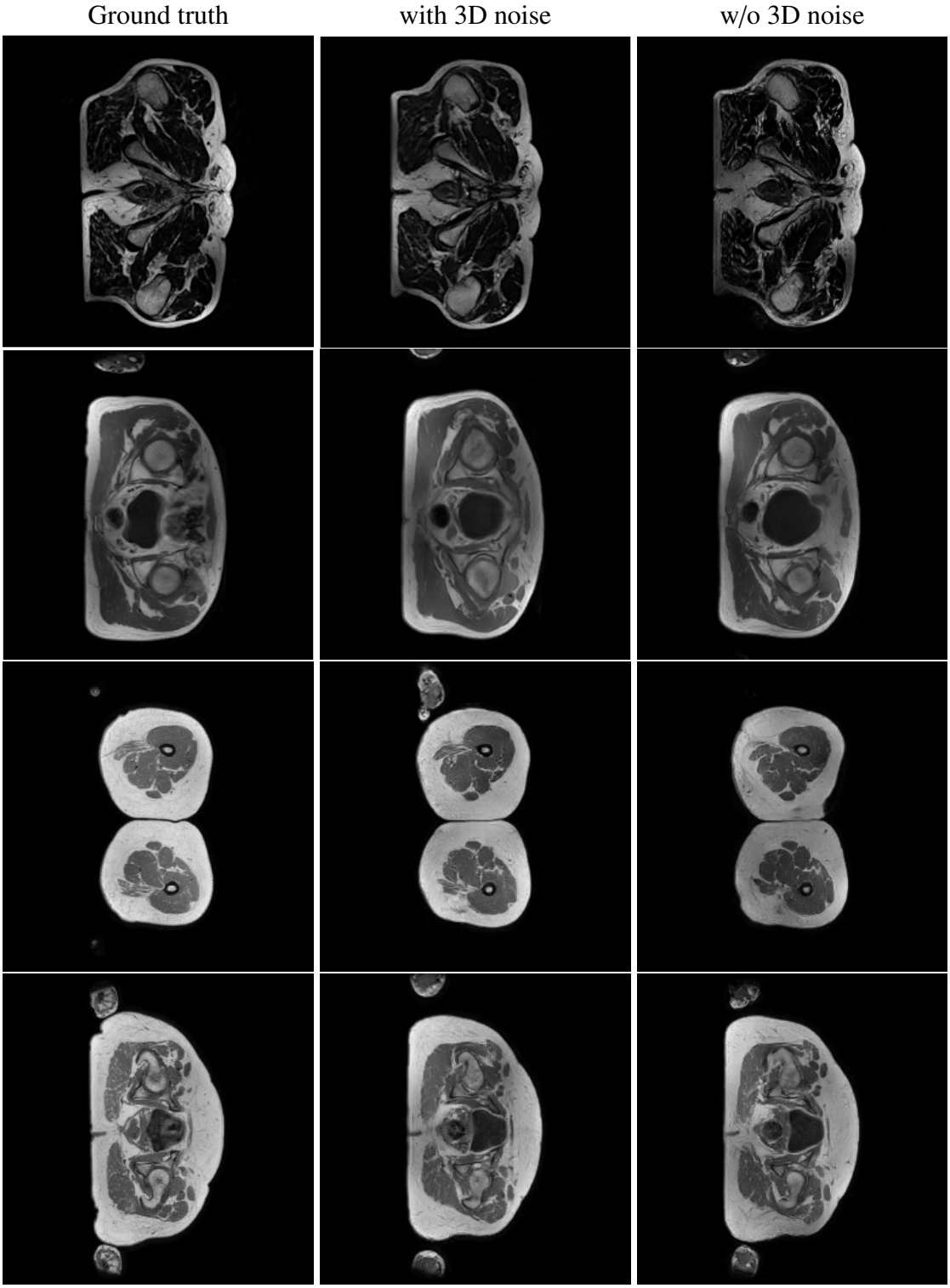

Figure A.5.: Qualitative comparison of Wavelet generator with and without 3D noise.

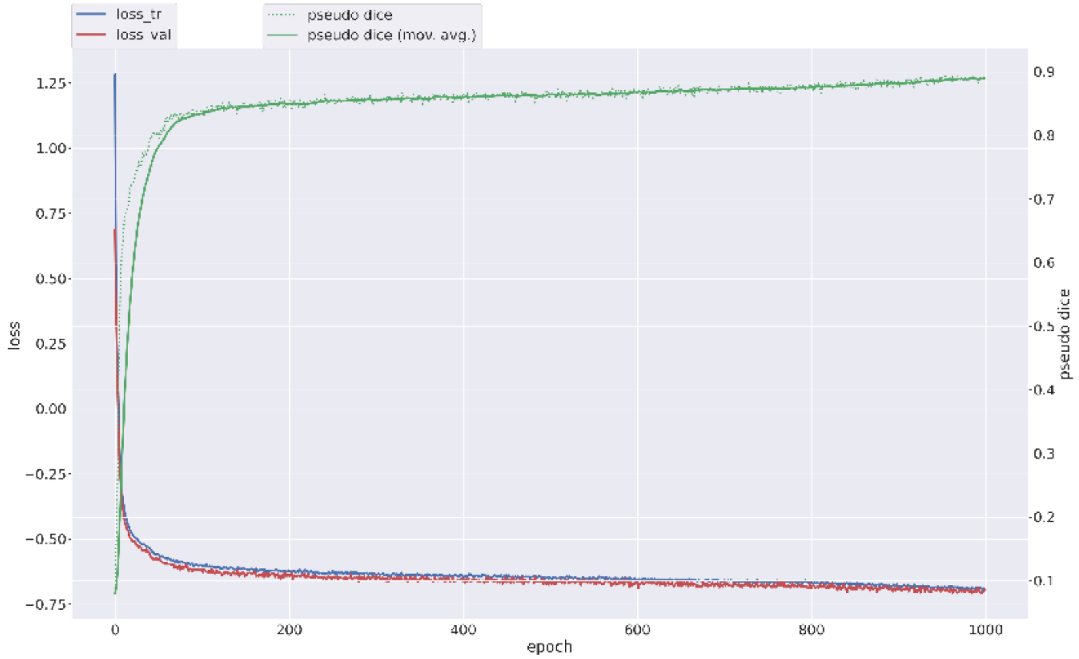

Figure A.6.: Training process of nnUNet

This image A.6 illustrates the training progression of an nnU-Net model across epochs. The blue and red lines represent the training and validation losses, respectively, both of which decline as the model learns, indicating improvement in prediction accuracy. The green dotted line depicts the Dice coefficient, a performance metric for segmentation tasks, which stabilizes after initial training, suggesting the model's consistent performance. The plateauing of loss and stability of the Dice score imply that the model is reaching its optimal performance as training progresses.

A.2. More tables

Table A.1.: CT classes of AutoPET dataset

ID	Percent/%	Class ID	Percent/%	Class ID	Percent/%
0	**81.4155**	13	0.0016	26	0.2103
1	0.0025	14	0.0252	27	0.4839
2	0.0827	15	0.2208	28	0.0245
3	0.0425	16	0.0073	29	0.0180
4	0.3732	17	0.1037	30	0.0397
5	0.2138	18	0.0020	31	0.0014
6	0.1233	19	0.0063	32	0.0062
7	0.6116	20	0.0025	33	**5.7686**
8	0.0214	21	0.2879	34	**6.5854**
9	0.0907	22	0.0009	35	1.6175
10	0.0067	23	0.0098	36	0.1043
11	0.0286	24	0.1551		
12	1.2245	25	0.0803		

Table A.1 presents the distribution of CT classes in the AutoPET dataset. It lists various classes (Class ID) along with their corresponding percentages (Percent/%) in the dataset. Each class is assigned a unique ID and is accompanied by the frequency of its occurrence within the entire dataset. Certain classes (like Class ID 0, 33, and 34) occupy a relatively high proportion in the dataset, with Class ID 0 (background) having the highest percentage at 81.4155%. On the other hand, some classes (such as Class ID 1, 13, and 31) have a very low frequency of occurrence. This distribution may reflect the class unbalance problem of the dataset.

Table A.2.: IoUs of generated CT images

ID	IoU	Class ID	IoU	Class ID	IoU
0	**0.9786**	13	0.0079	26	0.1156
1	0.0113	14	0.0498	27	0.4564
2	0.2423	15	0.1935	28	0.0927
3	0.0222	16	0.1096	29	0.0346
4	0.0687	17	0.1175	30	0.1238
5	0.0610	18	0.0004	31	0.0014
6	0.0940	19	0.0228	32	0.0133
7	0.1099	20	0.0120	33	**0.7902**
8	0.0148	21	0.1293	34	**0.7665**
9	0.0506	22	0.0001	35	0.2965
10	0.0178	23	0.0164	36	0.1356
11	0.1586	24	0.1851		
12	0.0936	25	0.1171		

Table A.3.: Segmented CT classes of SynthRAD2023 [1] dataset

ID	Class	ID	Class	ID	Class	ID	Class
0	Background	8	Lung	16	Colon	24	Autochthon
1	Kidney	9	Vertebrae	17	Ribs	25	Iliopsoas
2	Vessels	10	Esophagus	18	Humerus	26	Urinary Bladder
3	Gallbladder	11	Trachea	19	Scapula	27	Skin
4	Liver	12	Heart	20	Clavicula	28	Spleen
5	Stomach	13	Pulmonary Artery	21	Femur	29	Fat
6	Pancreas	14	Small Bowel	22	Hips	30	Skeletal Muscle
7	Adrenal	15	Duodenum	23	Sacrum		

Table A.3 presents a comprehensive listing of segmented CT classes within the SynthRAD2023 dataset [1]. This table categorizes various anatomical structures, each assigned with a unique identifier (ID) and a corresponding class name. The classes range from common anatomical features such as 'Kidney' and 'Liver' to more specific structures like 'Pulmonary Artery' and 'Iliopsoas'. The dataset's segmentation extends to a total of 30 distinct classes, including but not limited to organs, bones, and tissues.

List of Figures

1.1. A pair of corresponding MR (left) and CT (right) images and their difference 1
1.2. Example of paired data and unpaired data 2

2.1. (a) GAN architecture (b) Conditional GAN (CGAN) architecture 6
2.2. Our task described as a domain adaptation problem 7
2.3. USIS model architecture [8] 10

3.1. Overview of our proposed Med-USIS model 14
3.2. SPADE structure [21] 15
3.3. Structure of the U-Net based Segmentator 17
3.4. Outputs of the U-Net 18
3.5. Outputs of the U-Net 19
3.6. Overview of pre-processing 20
3.7. Visualization of the process of removing background artifact 21

4.1. Example of CT image from AutoPET 23
4.2. Overview of 117 anatomical structures segmented by the TotalSegmentator [6] 24
4.3. MR generation results and MAE 29
4.4. Comparison of our methods using MAE 30
4.5. Output of the U-Net 31
4.6. Qualitative comparison of our two generators. 32
4.7. Failure cases 33
4.8. Normal vs Truncated Gaussian distribution 34
4.9. Variation of FID and PSNR with "a" value 35

A.1. CT images from supervised model at very beginning. 39
A.2. CT images from supervised model with mask loss and artifacts removement. . 40
A.3. MR images from unsupervised model with Wavelet generator with mask loss. 41
A.4. Comparison of Wavelet and OASIS generator with and without mask loss. . . 42
A.5. Qualitative comparison of Wavelet generator with and without 3D noise. . . . 43
A.6. Training process of nnUNet 44

List of Tables

3.1. Parameter count of model components 13

4.1. Ablation study on AutoPET 28
4.2. Ablation study on SynthRAD2023 29
4.3. Ablation on sample interval 35
4.4. Ablation on 3D noise 36

A.1. CT classes of AutoPET dataset 45
A.2. IoUs of generated CT images 45
A.3. Segmented CT classes of SynthRAD2023 [1] dataset 46

Bibliography

[1] A. Thummerer, E. van der Bijl, A. Galapon, J. J. C. Verhoeff, J. A. Langendijk, S. Both, C. N. A. T. van den Berg and M. Maspero, "Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy," *Medical Physics*, vol. 50, no. 7, p. 4664–4674, Jun. 2023. [Online]. Available: http://dx.doi.org/10.1002/mp.16529

[2] N. A. Office of the Federal Register and R. Administration., "Health insurance portability and accountability act of 1996," pp. 104–191, August 21, 1996. [Online]. Available: https://www.govinfo.gov/app/details/PLAW-104publ191

[3] F. S. Zadeh, S. Molani, M. Orouskhani, M. Rezaei, M. Shafiei and H. Abbasi, "Generative adversarial networks for brain images synthesis: A review," 2023.

[4] X. Yi, E. Walia and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, Dec. 2019. [Online]. Available: http://dx.doi.org/10.1016/j.media.2019.101552

[5] C.-B. Jin, H. Kim, W. Jung, S. Joo, E. Park, A. Y. Saem, I. H. Han, J. I. Lee and X. Cui, "Deep ct to mr synthesis using paired and unpaired data," 2018.

[6] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach and M. Segeroth, "Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1148/ryai.230024

[7] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert and K. H. Maier-Hein, "nnu-net: Self-adapting framework for u-net-based medical image segmentation," 2018.

[8] G. Eskandar, M. Abdelsamad, K. Armanious and B. Yang, "Usis: Unsupervised semantic image synthesis," *Computers & Graphics*, vol. 111, pp. 14–23, 2023.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[10] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[11] A. Farahani, S. Voghoei, K. Rasheed and H. R. Arabnia, "A brief review of domain adaptation," 2020.

[12] X. Liu, C. Yoo, F. Xing, H. Oh, G. E. Fakhri, J.-W. Kang and J. Woo, "Deep unsupervised domain adaptation: A review of recent advances and perspectives," 2022.

[13] R. Zhang, T. Pfister and J. Li, "Harmonic unpaired image-to-image translation," 2019.

[14] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang and R. Webb, "Learning from simulated and unsupervised images through adversarial training," 2017.

[15] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," 2018.

[16] Y. Chen, G. Li, C. Jin, S. Liu and T. Li, "Ssd-gan: Measuring the realness in the spatial and spectral domains," 2020.

[17] J. Wang, X. Deng, M. Xu, C. Chen and Y. Song, "Multi-level wavelet-based generative adversarial network for perceptual quality enhancement of compressed video," 2020.

[18] H. Huang, R. He, Z. Sun and T. Tan, "Wavelet domain generative adversarial network for multi-scale face hallucination," *International Journal of Computer Vision*, vol. 127, pp. 763 – 784, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:60441776

[19] Y. Liu, Q. Li and Z. Sun, "Attribute-aware face aging with wavelet-based generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[20] R. Gal, D. Cohen, A. Bermano and D. Cohen-Or, "Swagan: A style-based wavelet-driven generative model," 2021.

[21] T. Park, M. Liu, T. Wang and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," *CoRR*, vol. abs/1903.07291, 2019. [Online]. Available: http://arxiv.org/abs/1903.07291

[22] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen and T. Aila, "Analyzing and improving the image quality of stylegan," 2020.

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

[24] B. Zhou, H. Zhao, X. Puig, T. Xiao, S. Fidler, A. Barriuso and A. Torralba, "Semantic understanding of scenes through the ade20k dataset," 2018.

[25] H. Caesar, J. Uijlings and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," 2018.

[26] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," 2015.

[27] V. Sushko, E. Schönfeld, D. Zhang, J. Gall, B. Schiele and A. Khoreva, "You only need adversarial supervision for semantic image synthesis," 2021.

[28] A. Dosovitskiy and T. Brox, "Generating images with perceptual similarity metrics based on deep networks," 2016.

[29] E. Schönfeld, B. Schiele and A. Khoreva, "A u-net based discriminator for generative adversarial networks," 2021.

[30] S. Gatidis, T. Hepp, M. Früh, C. La Fougère, K. Nikolaou, C. Pfannenberg, B. Schölkopf, T. Küstner, C. Cyran and D. Rubin, "A whole-body fdg-pet/ct dataset with manually annotated tumor lesions," *Scientific Data*, vol. 9, no. 1, p. 601, 2022.

[31] R. Zhang, P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[33] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

Declaration

Herewith, I declare that I have developed and written the enclosed thesis entirely by myself and that I have not used sources or means except those declared.

This thesis has not been submitted to any other authority to achieve an academic grading and has not been published elsewhere.

Stuttgart, TBD Date of sign. Wenwu Tang