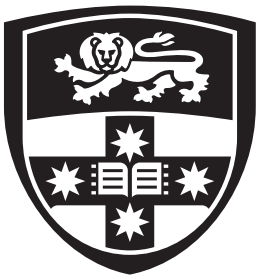
Standard-dose PET(SPET) Images Synthesizing

Boyuan Tan Mphil



THE UNIVERSITY OF
SYDNEY

A thesis submitted in fulfillment of the requirements
of the degree of Master of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney

August 2025

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

During the preparation of this thesis, ChatGPT was used for the purposes of text enhancement, including improving sentence structure, clarifying phrasing, and paraphrasing content for better readability. Where any text was modified by generative AI, the author then reviewed the resulting content for any errors, inaccuracies or biases, and modified it as required.

The author takes full responsibility for the submitted thesis and ensures the work is their own and has used generative AI within the parameters of use, see University of Sydney generative AI guide for researchers .

Authorship Attribution Statement

Chapter 3 of this thesis is based on a publication B. Tan, Y. Xue, L. Bi, and J. Kim, "A reverse method of data augmentation for high quality pet image synthesis," in 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2024, pp. 427–434. I designed the method, conducted the experiments, analysed the data and drafted the manuscript.

Chapter 4 of this thesis is based on a publication the following two publication: B. Tan, Y. Xue, L. Bi, and J. Kim, "Full-trsun: A full-resolution transformer unet for high quality pet image synthesis," in International Workshop on Machine Learning in Medical Imaging. Springer, 2024, pp. 238–247, and B. Tan, Y. Xue, L. Bi, and J. Kim, "TCCA-Net: A Time-Controlled Channel-Spatial Attention Enhanced ResUNet to help with the PET image synthesizes,", in preparation. I designed the method, conducted the experiments, analysed the data and drafted the manuscript.

In addition to the statements above, in cases where I am not the corresponding author of a published item, permission to include the published material has been granted by the corresponding author.

Abstract

Positron emission tomography (PET) is a functional imaging modality that uses a radioactive tracer to visualise and quantify metabolic processes in the human body. It has demonstrated considerable clinical value in oncology, neuropsychiatry, and cardiology, showing substantial promise for advancing cancer diagnosis and management, cardiac care and surgery, as well as neurological and psychiatric applications.

[18F] fluorodeoxyglucose (FDG) is the most commonly used tracer for PET imaging. After injecting the radio tracer into the patient, tiny particles released from the tracer interact with the body to create energy signals that the PET scanner detects; A PET image is then generated by applying a reconstruction algorithm to turn the signals into an image. However, repeated imaging poses health risks due to cumulative radiation exposure. To address this critical issue, there have been attempts to lower the injected activity to get low-dose PET (LPET) images. Because LPET involves less radiotracer accumulation than standard-dose PET (SPET), the reconstructions are noisier and can compromise diagnostic utility. To restore image quality, recent work employs deep-learning methods to denoise LPET and synthesise images that closely approximate SPET quality.

In recent years, there has been remarkable progress in LPET enhancement through the use of deep learning based 'image synthesis'. Image synthesis refers to the process of generating high-quality images from lower-quality or incomplete data, and it has become the state-of-the-art in LPET enhancement. Despite these advances, deep learning-based image synthesis methods still face challenges, such as availability of diverse datasets, generalization across diverse datasets, robustness against noise and artifacts, and ensuring clinical reliability and interpretability of synthesized images.

The objectives of this thesis are to investigate novel strategies and advanced deep learning methods for PET image synthesis tasks. Firstly, SPET images are widely available as they are routinely acquired, whereas LPET images are relatively scarce as they are primarily for research purpose. Therefore, paired low-to-high-dose PET image datasets are challenging to obtain. However, deep learning methods require extensive data for training. To address this, in this thesis, a new data augmentation strategy is proposed in which synthetic LPET images were generated from existing SPET images and used to augment the model training. This novel approach improved model performance by increasing the diversity and quantity of training data and enhanced the model's ability to synthesize SPET from LPET.

Secondly, LPET images are acquired at various dose levels, such as 50%, 10%, 1%, etc. Training models with LPET images from these different dose levels significantly increase model training time and the need for extensive training datasets that cover all the different dose levels. This limits the generalizability of existing methods, as they are often optimized for specific dose levels and struggle to adapt to unseen or varying dose conditions. To address this challenge, a time-controlled deep learning method is proposed that is capable of improving image quality while also demonstrating greater generalizability across LPET images with unseen noise levels. Here, the timestep was set to represent different noise levels. The proposed model was trained on the lowest dose level and tested across various dose levels. Unlike previous approaches, this model achieved robust performance even with images with noise characteristics it has not seen during training, highlighting its flexibility and broader applicability in clinical scenarios.

Both methods were trained on a public dataset called ultra-low-to-high PET dataset. A subset containing PET images scanned by a Siemens scanner was selected, with each subject having seven different dose levels. The data was acquired in 'list mode,' allowing images to be reconstructed at different time points to simulate various low-dose PET conditions.

The results of the data augmentation method demonstrate that the proposed training strategy enhanced performance compared to using a smaller dataset, especially when integrated with various state-of-the-art (SOTA) PET synthesize methods. The second proposed method's outperformed the comparison SOTA, with the incorporation of the time-controlled strategy contributing to the performance, particularly when testing on dose-level PET images not included in the training dataset.

Acknowledgements

I would like to acknowledge that my MPhil journey has been a challenging yet rewarding experience. Throughout this time, I have learned and grown, and I am truly grateful to those who have supported me along the way.

First, I would like to express my sincere gratitude to my supervisor, Professor Jinman Kim, for his guidance, encouragement, and patience throughout my MPhil journey. His expertise and insightful advice have been invaluable in shaping this research, from refining ideas to overcoming technical challenges. His continuous support and constructive feedback have not only helped me improve my research skills but also deepened my understanding of PET image synthesis and deep learning applications. I truly appreciate the time and effort he have dedicated to mentoring me, and I am grateful for his belief in my abilities, especially during the more difficult phases of my research.

Secondly, I extend my sincere appreciation to Associate Professor Lei Bi for his valuable support, insightful contributions, and significant help with my paper writing. His expert guidance and consistent encouragement played a crucial role in improving the quality of my research and navigating complex technical aspects.

I would also like to extend my gratitude to Doctor Yuxin Xue for her invaluable support, guidance, and encouragement throughout my research. Her expertise and willingness to share her knowledge have greatly helped me navigate various challenges, from understanding complex methodologies to improving my experimental design. She has always been generous with her time, offering constructive feedback and insightful discussions that have significantly contributed to the progress of this thesis. Beyond academic support, her mentorship and encouragement have provided me with confidence and motivation, making my research journey both enriching and fulfilling.

I would also like to appreciate Yuan Yuan, Doctor Mingyuan Meng, Yongpei Ma, Shuchang Ye, Mingxiao Tu, Yue Xia, Zhuyi Lu, Jiadi Dong, Xiaoshuang Li, Pengyu Wang and other colleagues in the Biomedical Data Analysis and Visualisation (BDAV) research group. Their support has played an important role in my MPhil study.

I am deeply grateful to my friends for their unwavering support and encouragement throughout my MPhil journey. Special thanks to Xiang Li, Lingyu Wang, and Yaze

Wu—their belief in me has been a constant source of motivation, especially during times of doubt and difficulty. I truly appreciate their companionship and the comfort they offered during stressful moments in life.

Finally, I would like to express my deepest gratitude to my parents. Their unconditional love, endless patience, and constant encouragement have been the foundation of my journey. They have always believed in me, even when I doubted myself, and their words of comfort and reassurance have given me the courage to face every challenge. Their selfless support—both emotionally and practically—has allowed me to focus on my studies without worry. From the smallest acts of care to the biggest sacrifices made quietly behind the scenes, they have been my strongest pillars. I owe everything to their unwavering faith in me, and I am forever thankful for the strength, resilience, and values they have instilled in me.

This thesis is not just the result of my own efforts but also a reflection of the support, kindness, and belief of those who have stood by me every step of the way. I am forever grateful for their presence in my life.

List of Publications

Published or Accepted:

1. **B. Tan**, Y. Xue, L. Bi, and J. Kim, "Full-trsun: A full-resolution transformer unet for high quality pet image synthesis," in International Workshop on Machine Learning in Medical Imaging. Springer, 2024, pp. 238–247.

2. **B. Tan**, Y. Xue, L. Bi, and J. Kim, "A reverse method of data augmentation for high quality pet image synthesis," in 2024 International Conference on Digital Image Computing: Techniques and Applications (DICTA). IEEE, 2024, pp. 427–434.

Under Preparation, Review or Revision:

1. **B. Tan**, Y. Xue, L. Bi, and J. Kim, "TCCA-Net: A Time-Controlled Channel-Spatial Attention Enhanced ResUNet to help with the PET image synthesizing," (Journal Paper, In Preparation)

Contents

List of Figures

List of Tables

List of Acronyms

CT	Computerized Tomography
MRI	Magnetic Resonance Imaging
PET	Positron emission tomography
FDG	[18F] fluorodeoxyglucose
LPET	low-dose PET
SNR	signal-to-noise ratio
SPET	standard-dose PET
GLUT	glucose transporters
SPECT	single photon emission computed tomography
FCN	fully convolutional network
GAN	generative adversarial network
AE-ResUNet	Attention-Enhanced ResUNet
CAE	channel-wise attention encoder
RAE	ResNet-Attention Encoder
SiLU	Sigmoid-Weighted Linear Unit
DRF	dose reduction factor
PSNR	Peak Signal-to-Noise Ratio
SSIM	Structural Similarity Index
MSE	Mean Squared Error
NRMSE	Normalized Root Mean Squared Error
SOTA	state-of-the-art
TOF	Time-of-Flight
OSEM	Ordered Subset Expectation Maximization
XCA	Cross-Covariance Attention
ALARA	"As Low As Reasonably Achievable"
ICRP	International Commission on Radiologic Protection
CNN	Convolutional Neural Network
dNet	dilated convolutional neural network
DLE	deep learning enhancement
3D c-GAN	3D Conditional GAN
SS-AEGAN	Self-Supervised Adaptive Residual Estimation Generative Adversarial Network

PT-WGAN	Parameter-Transferred Wasserstein GAN
SUVs	standardized uptake values
BiC-GAN	Bidirectional Contrastive GAN
MCI	mild cognitive impairment
3D CVT-GAN	3D Convolutional Vision Transformer GAN
GFP	global frequency parser
RAT	Region Attention Transformer
PK-TriDo	Prior Knowledge-guided Triple-Domain Transformer-GAN
DDPMs	Denoising Diffusion Probabilistic Models
DTM	Diffusion Transformer Model
RN	Region-Adaptive Normalization

Chapter 1

Introduction

1.1 Background and Motivation

Medical images, acquired from different imaging modalities, depicts the human body's structure, function and pathology, providing detailed information for clinical and research applications [5]. There are a variety of medical imaging modalities such as X-rays, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US) and Positron emission tomography (PET). The focus of this thesis is with PET which is a functional imaging modality that uses radioactive tracers to visualize and quantify metabolic processes and other physiological activities [48]. Various tracers can be used for PET imaging, with [18F] fluorodeoxyglucose (FDG) being the most commonly used for cancer detection due to its effectiveness in identifying increased glucose metabolism in malignant cells [82]. The clinical value of PET has been widely recognized in oncology, neuropsychiatry, and cardiology, demonstrating significant potential for future applications in cancer diagnosis and management, cardiology and cardiac surgery, as well as neurology and psychiatry [48].

PET imaging process [40], shown in Figure 1.1 involves the administration of a radiopharmaceutical, typically FDG, which is taken up by metabolically active tissues. The emitted positrons from the radiotracer interact with electrons in the body, resulting in the release of two 511 keV photons that are detected by PET scanners.

These detectors use scintillation crystals to capture coincidence photons, enabling the identification of regions with abnormal metabolic activity. Finally, the PET image is obtained through the application of appropriate reconstruction techniques and algorithms.

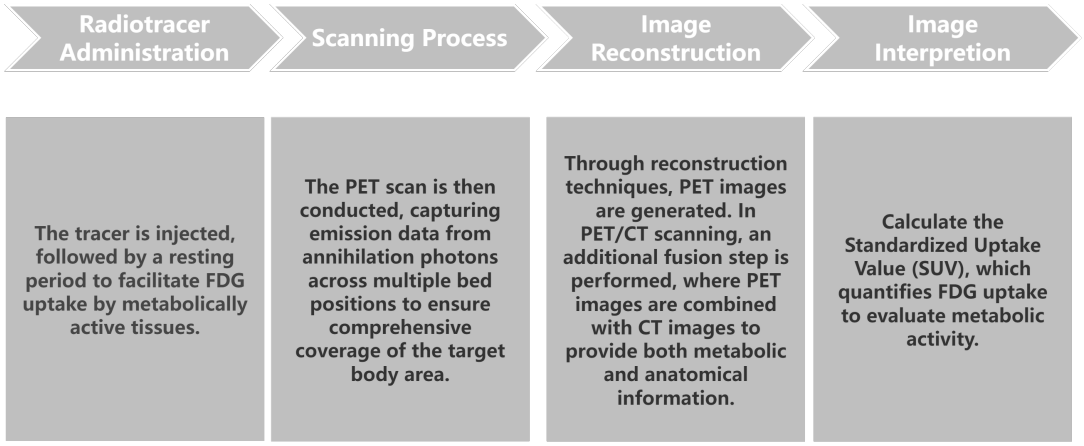


Figure 1.1 – PET imaging Process

Despite the proven clinical value and wide range of clinical applications, there are growing concerns about the potential health risks of cumulative radiation exposure from repeated PET imaging. Oncology patients, in particular, often undergo frequent PET imaging, which has been shown to increase their long-term risk of radiation-induced malignancies [76]. To mitigate the risk from repeated exposure, one approach is to reduce the injected dose, leading to low-dose PET (LPET) scans [39]. Compared to standard-dose PET (SPET), LPET images have lower quality because shorter scans introduce more noise, resulting in a reduced signal-to-noise ratio (SNR), often due to reduced radiotracer dose . Quantitatively, they are measured using metrics such as peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). Qualitatively, they show blurring, artifact presence, and impaired lesion visibility, which may limit clinical interpretability. However, at lower noise levels associated with reduced dose, the resulting image quality degrades significantly, making LPET impractical for diagnostic use. Deep learning methods can effectively learn patterns

from pairs of low-quality and high-quality images. This enables them to reduce noise and recover essential image details from noisy LPET scans, generating clearer, SPET-like images and potentially improving diagnostic accuracy. For instance, a PET image that clearly reveals hypermetabolic lesions or preserves quantification (e.g., SUV) for treatment planning is of greater value than one that is merely noise-free. Therefore, improving image quality should aim not just at aesthetic restoration, but at retaining diagnostically relevant information critical for clinical decision-making.

In recent years, image synthesis methods, which involve generating high-quality images from lower-quality or incomplete data, have been widely applied in the synthesis of standard-dose PET (SPET) images from low-dose PET (LPET) images using deep learning techniques. Figure 1.2 visualizes PET images of the same patient at different dose levels. In this context, DRF stands for Dose Reduction Factor. A **DRF100** means the dose is reduced by a factor of 100. This implies that the injected tracer is only $\frac{1}{100}$ of the full (standard) dose, which equals 1%. Similarly, **DRF50** means the dose is reduced by a factor of 50, resulting in $\frac{1}{50} = 2\%$. And **Full_dose** represents an injection with the standard dose tracer. As the DRF value increases (meaning the injected tracer dose decreases), images become noisier, and signals become less reliable. For instance, at higher DRF levels such as DRF100 and DRF50, the images show higher noise compared to the Full_dose, making signals in organs like the liver appear falsely elevated. Conversely, genuine signals, such as the high uptake typically seen in the right kidney, become unclear or completely hidden due to this noise. Such kidney uptake usually starts to appear more clearly at lower DRF values like DRF20, becoming fully distinguishable only at very low DRFs, such as DRF2, where the noise is substantially reduced.

The first gap in PET image synthesis to building image synthesis models is the availability and quality of PET datasets. Although SPET images are more accessible, real LPET images remain scarce. Paired low-dose and high-dose PET datasets are particularly difficult to obtain due to ethical and practical constraints related to radiation exposure. It is ethically problematic because exposing patients to both low- and high-dose PET scans increases radiation risk without clinical benefit. Practi-

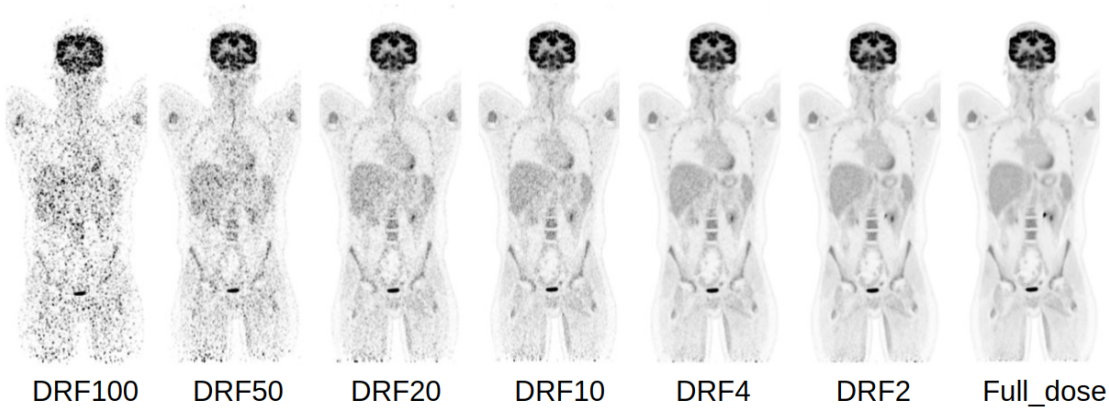


Figure 1.2 – Different dose level PET images

cally, performing two scans is costly, time-consuming, and burdensome for patients and healthcare systems.

The second gap in PET image synthesis lies in the limited generalizability of current models. While prior studies focus on capturing more information to improve image quality, these models are typically trained on specific dose-level data and test on specific dose-level, making them less effective when applied to PET images with unseen or varying dose levels in real-world settings. This reduces their effectiveness in real-world scenarios where PET images can have varying dose levels. Collecting training data for every possible noise or dose condition is constrained by factors like high costs, limited access to varied patient data, ethical restrictions on radiation exposure, and the extensive diversity of imaging protocols in clinical practice. Since it is impractical to collect training data covering all possible noise or dose conditions, models often fail when applied to real-world cases that deviate from their training distribution.

1.2 Contribution

This thesis presents two contributions designed to overcome the identified gaps (Figure 1.3) to the field of PET image synthesis, focusing on using deep learning and data

augmentation methods to improve the synthesis of higher quality SPET images from LPET counterpart.

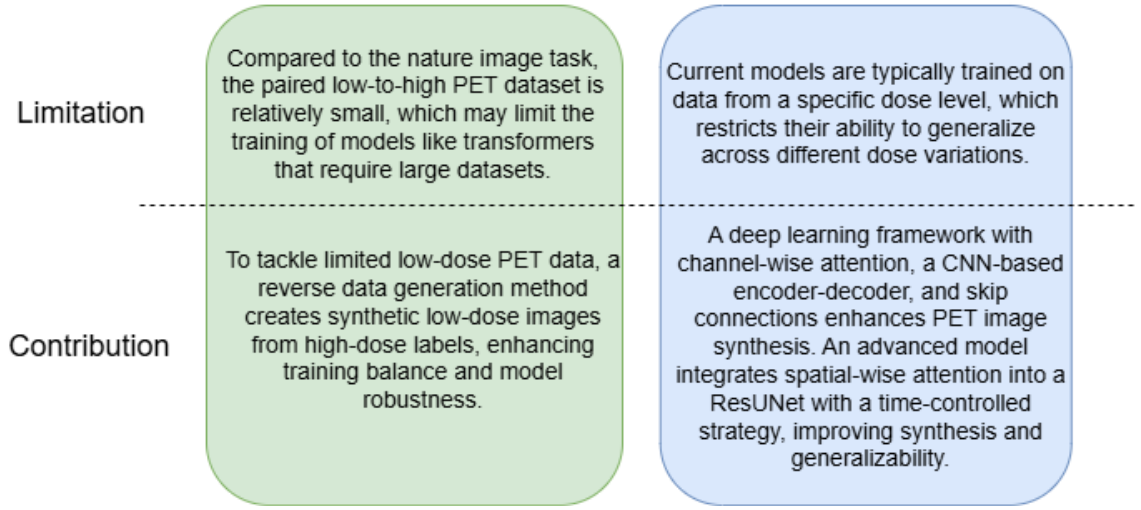


Figure 1.3 – Limitations and Contributions

1. **Proposed a data augmentation strategy for PET synthesis with imbalanced datasets:**

   A key challenge in PET imaging research is the scarcity of real LPET images, particularly paired datasets with corresponding SPET images. To address this, this thesis introduces a novel reverse data generation approach, which has not been explored in existing augmentation methods. Unlike traditional augmentation techniques that rely on large amounts of unlabeled input data to enhance image clarity or generate target data, our method synthesizes low-dose PET images from high-dose counterparts using a novel deep learning-based transformation approach. This reverse approach significantly expands the availability of LPET data, enabling more effective model training while preserving realistic noise distributions. By integrating these synthetic LPET images into the training pipeline, data imbalance was mitigated and the model performance was enhanced in the synthesis of SPET tasks.

2. **Development of a time-controlled channel-spatial attention enhanced ResUNet to improve the model generalizability:**

The main contribution of this work is the introduction of a time-controlled strategy designed to simulate PET images with varying noise levels by introducing controlled degradation into the target images. Specifically, a timestep t is used to regulate the intensity of noise, guiding the adaptive synthesis process by incorporating t as an additional input and calculating loss against low-dose PET (LPET) images. To complement this strategy, a deep learning framework integrating channel-wise attention within a CNN-based encoder-decoder architecture with skip connections was developed. This framework enhances the model's ability to emphasize relevant feature channels, improving high-dose PET image synthesis from low-dose inputs. Additionally, spatial-wise attention was integrated into the final encoder block of a ResUNet-based structure to capture both local and global spatial dependencies. This refined architecture effectively captures channel-wise and spatial dependencies, resulting in improved synthesis performance and enhanced model generalizability.

1.3 Thesis Organization

The remainder of this thesis is organized as follows:

- Chapter 2 reviews related work, focusing on deep learning methods for various medical imaging tasks, with a particular emphasis on PET images and the synthesis of high-quality PET images from LPET images.
- Chapter 3 presents a data augmentation strategy for PET synthesis, assuming a large dataset containing SPET images and a subset with corresponding LPET images at varying dose levels, reflecting real clinical scenarios.
- Chapter 4 introduces a novel deep learning architecture that integrates channel-wise attention, ResNet blocks, and spatial attention. Additionally, a time-controlled strategy is incorporated to enable the model to adapt to PET images at different dose levels with varying noise.

- Chapter 5 explores the strengths, limitations, and potential future directions in this field.

Chapter 2

Literature Review

2.1 Introduction

PET imaging has become an indispensable imaging modality in oncology, neurology, and cardiology [9]. It is a nuclear medicine imaging modality that involves several key stages in its usage: administration of a radiotracer emitting positrons (such as FDG), detection of gamma rays emitted during positron annihilation, and reconstruction of these events into diagnostic images. Advances in PET imaging, including the transition from 2D to fully 3D scanning, the integration of iterative reconstruction algorithms, and improved attenuation correction methods—particularly through PET/CT integration—have greatly enhanced PET's diagnostic accuracy, image quality, and clinical usability [51]. During the imaging process, FDG is transported into cancer cells primarily via glucose transporters (GLUT), where it accumulates due to increased metabolic demand. This selective uptake provides a basis for visualizing tumors in PET imaging [30].

However, traditional PET imaging modalities are often limited by issues such as radiation exposure and prolonged scan times [74]. To address these challenges, the integration of artificial intelligence (AI), particularly deep learning methodologies, into PET imaging has gained strong attention. AI-driven approaches have demonstrated

considerable potential in enhancing image reconstruction quality, reducing noise, synthesizing high-quality PET images from lower-dose scans, and ultimately improving diagnostic accuracy and clinical outcomes [49]. Existing research approaches the recovery of full PET imaging data from noisy or incomplete data using various methods, including denoising, enhancement, and synthesis. While denoising aims to remove noise from acquired images, enhancement focuses on improving image quality, and synthesis seeks to generate high-quality images from significantly reduced radiation doses. All these methods focus on recovering high-quality PET images from low-quality inputs. In Section 2.2, various AI-driven approaches applied to PET image analysis will be discussed. In Section 2.3, we will specifically explore deep learning methods aimed at synthesizing high-quality PET images from low-quality or low-dose PET data.

2.2 AI in PET imaging

AI has been extensively applied across various domains, notably transforming medical imaging through significant enhancements in image quality and diagnostic accuracy. In PET imaging, AI-driven methods, particularly deep learning, have been widely employed for recovering tasks such as image denoising, enhancement, synthesis, as well as image translation and reconstruction [6, 47, 55, 71]. While denoising and enhancement approaches address the removal of noise artifacts and the improvement of image clarity, respectively, synthesis methods generate high-quality images from extremely low-dose inputs. In addition, AI-driven image translation converts data across different imaging modalities or conditions, and reconstruction algorithms directly enhance images from raw measurement data [28].

In 1991, Floyd first proposed an artificial neural network for reconstructing quantitative single photon emission computed tomography (SPECT) images [20]. Since then, numerous researchers have explored various deep learning models to further enhance image reconstruction quality. For instance, Häggström et al. introduced a fully convolutional network (FCN) architecture designed for direct PET image re-

construction tasks [26].Then Hashimoto introduced a novel ReconU-Net architecture, which uniquely integrates the physics-based back-projection operation into U-Net's skip connections for direct PET image reconstruction task, especially when trained on limited simulated data [29].

Deep learning-based image translation significantly enhance PET imaging by (1) supplementing incomplete or missing data, (2) reducing the required number of PET scans, and (3) augmenting datasets [49]. Ben-cohen proposed a model which combines a FCN with a conditional generative adversarial network (GAN) for the generation of virtual PET images from CT scans [6]. Apoorva et al. proposed a globally and locally aware generative adversarial network (GLA-GAN) to synthesize realistic FDG-PET images from MRI, significantly improving image quality and diagnostic performance in the detection of Alzheimer's disease [67]. Dac introduced a novel conditional diffusion model, named CPDM, for translation from CT to PET images [52].

Deep learning models have also been applied to PET image denoising, enhancement, and synthesis [12, 34, 62]. These methods share a similar goal, which is to improve image quality by removing noise, making images clearer, or generating high-quality images from low-quality scans. These methods will be introduced in more detail in Section 2.3.

2.3 Deep learning methods for PET image synthesizing

In recent years, many different neural network architectures have been developed to create high-quality PET images from low-quality or low-dose input data. To clearly understand and compare these approaches, this literature review categorizes existing studies based on the type of deep learning models employed. The models discussed include Convolutional Neural Network (CNN) [42], GAN [24], Transformer [75] and Diffusion Models [31], each demonstrating unique strengths and capabilities in syn-

thesizing high-quality PET images from low-quality inputs.

2.3.1 CNN based methods

Sustained progress has been achieved in synthesizing LPET images, with deep learning techniques emerging as key contributors to these advancements. CNN have become foundational in image synthesis techniques. The U-Net [59] and 3D U-Net [11], initially developed for biomedical image segmentation, have also been effectively applied to PET image synthesizing [64].

Building on the concept of residual learning [92], Yan-Ran combined the residual block with UNet to generate diagnostic PET image from ultra-low-dose PET images [80]. Also utilized residual U-Net, Sano et al. [63] focused on range verification. Their method was tested on simulated and experimental data, demonstrating that the Residual U-Net preserved peak signal regions better than conventional filters, providing accurate range estimation for dose verification and potentially reducing PET measurement times in clinical applications. Spuhler et al. [69] introduced a novel dilated convolutional neural network (dNet) to recover high-quality, full-count PET images from low-count data. This approach utilizes dilated convolutions to preserve image resolution and capture larger features without downsampling, achieving enhanced image quality over traditional U-Net and Gaussian filtering methods. Mehranian et al. [50] introduced a deep learning enhancement (DLE) model to improve whole-body PET image quality, reducing noise in short-duration scans. This model significantly improved image quality for oncology applications, allowing scan time or injected dose reduction while preserving diagnostic accuracy, emphasizing the viability of DLE in clinical PET applications.

However, CNN-based methods often struggle with capturing long-range dependencies, as convolutions inherently operate within a limited receptive field. Additionally, CNNs may oversmooth fine details, leading to potential loss of clinically significant information [68, 97].

2.3.2 GAN based methods

As GAN [24] become more and more popular recently, Wang et al. [79] implemented a 3D Conditional GAN (3D c-GAN)with a U-Net-like generator for high-quality SPET synthesis. The model's skip connections and concatenated layers facilitated detailed image reconstruction, while a unique loss function combining L1 and adversarial components further improved quality by reducing artifacts.

To address cross-tracer and cross-scanner variability, Xue et al. [85] developed an adaptive model using multi-stage GANs. The Self-Supervised Adaptive Residual Estimation Generative Adversarial Network (SS-AEGAN) dynamically estimated residual information, which helped correct discrepancies between synthesized and actual SPET images. The model's self-supervised pretraining enhanced its adaptability across different PET scan environments.

Gong et al. [23] proposed a Parameter-Transferred Wasserstein GAN (PT-WGAN) that reduced computational demands and improved training efficiency using parameter transfer. By leveraging both 2D and 3D convolutional layers, the model efficiently captured contextual information while maintaining structural fidelity in LPET reconstructions. CycleGAN Models with Supervised Learning have demonstrated success in PET denoising. Zhou et al. [96] and Zhao et al. [95] used CycleGAN architectures, adding supervised loss functions to enhance LPET recovery. Their models integrated cycle-consistency loss with a Wasserstein distance term, leading to more accurate SPET estimations and preserving standardized uptake values (SUVs), crucial for diagnostic consistency. Ouyang et al. [54] introduced a GAN framework that incorporated feature matching and task-specific perceptual loss for ultra-low-dose PET reconstruction. Their approach involved a U-Net-based generator and a multi-slice input strategy that significantly enhanced the network's ability to denoise and maintain structural integrity in PET images. This method demonstrated superior performance in visual quality metrics and diagnostic consistency compared to traditional methods.

Classification-Guided Approaches further optimize SPET synthesis by incorporat-

ing task-driven constraints. Fei et al. [19] proposed a Bidirectional Contrastive GAN (BiC-GAN), which utilized contrastive learning for domain alignment between LPET and SPET images. This design maximized shared information across domains, maintaining diagnostic details critical for conditions like mild cognitive impairment (MCI) detection. Luo et al. [46] introduced AR-GAN, which incorporates adaptive rectification to refine SPET synthesis outputs and spectral regularization to preserve high-frequency details, ensuring alignment in the frequency domain. Xue et al. [86] proposed a super-resolution refinement, CG-3DSRGAN, for PET image synthesis. This model incorporated a classification module to improve the accuracy of the low-to-high-dose translation, integrating a multi-tasking generator and a secondary network for spatial detail recovery. This architecture demonstrated superior performance across dose reduction levels, producing refined images with higher diagnostic value.

Despite their advantages, GANs face notable challenges, particularly in terms of training stability and output diversity. One common issue is mode collapse, where the generator learns to produce only a limited set of outputs, ignoring other modes in the data distribution. This significantly reduces the diversity and generalizability of the generated images, limiting their effectiveness in real-world applications [81].

2.3.3 Transformer based methods

Transformers [75], originally designed for natural language processing tasks, have shown significant potential in computer vision due to their ability to model long-range dependencies. Luo et al. [45] proposed the Transformer-GAN, combining CNNs for local feature extraction with a Transformer for global context modeling, achieving superior performance in SPET reconstruction compared to traditional GANs. Zeng et al. [91] introduced a 3D Convolutional Vision Transformer GAN (3D CVT-GAN), which integrated convolutional embeddings into the Transformer blocks to preserve local spatial details while capturing global semantic information for PET reconstruction.

TriDo-Former [15] represents a significant advancement in Transformer-based reconstruction by integrating triple-domain knowledge (sinogram, image, and frequency). The model employs two cascaded Transformers: SE-Former for denoising LPET sinograms while preserving their structure, and SSR-Former for reconstructing high-quality PET images using a global frequency parser (GFP) to retain high-frequency details. The TriDoRNet [38] and PK-TriDo [16] models further expanded on this. TriDoRNet reconstructed SPET from LPET by integrating information across projection, image, and frequency domains, with specialized networks for denoising and frequency adjustment to enhance high-frequency detail recovery. The Prior Knowledge-guided Triple-Domain Transformer-GAN (PK-TriDo) utilized prior domain knowledge and adaptive frequency parsing to improve reconstruction accuracy, particularly for ill-posed sinogram-to-image transformations.

Recent works have explored region-based attention mechanisms for medical image restoration. The Region Attention Transformer (RAT) [88] dynamically partitions images into semantic regions for targeted attention, reducing interference from irrelevant areas while enhancing detail in high-difficulty regions such as edges or lesions. This approach has proven effective in PET image synthesis, allowing for focused reconstruction of critical areas without losing overall context. Additionally, Li et al. proposed the PETformer [43], a U-Net-based architecture incorporating multi-headed attention blocks for short- and long-range dependency modeling. PETformer demonstrated significant improvements in denoising ultra-low-dose PET images by leveraging attention mechanisms to maintain anatomical details and signal-to-noise ratios across varying dose reductions.

However, Transformers require large-scale datasets for effective training, which is often impractical in medical imaging due to data scarcity. Additionally, their computational complexity is significantly higher on long sequences [25].

2.3.4 Diffusion based methods

Diffusion models have become increasingly popular due to their enhanced training stability and ability to generate high-quality images.. Ho et al. [31] introduced Denoising Diffusion Probabilistic Models (DDPMs), which add Gaussian noise to images through a forward diffusion process and learn to reverse it in a step-by-step manner. DDPMs have shown superior performance over GANs in producing high-quality medical images, including PET images.

Pan et al. [56] extended the application of DDPMs by proposing a high-efficiency denoising diffusion model (PET-CM) for whole-body PET reconstruction. The model significantly reduced the computational requirements while achieving high quantitative accuracy, demonstrating potential in clinical scenarios where quick image synthesis is essential. Similarly, Sanaat et al. [61] applied DDPMs to improve brain PET imaging, reducing radiation dose to as low as 5% of the full dose while maintaining diagnostic quality.

Transformer models, known for their ability to capture long-range dependencies, have been integrated with diffusion techniques for enhanced PET reconstruction. Huang et al. [33] proposed a Diffusion Transformer Model (DTM) that combines the powerful distribution mapping abilities of diffusion models with the Transformers' capability to capture spatial features. This model, guided by a joint compact prior, effectively denoises PET images while preserving critical diagnostic information.

Another significant advancement is the High-Frequency-guided Residual Diffusion Model (HF-ResDiff) [73], designed for multi-dose PET reconstruction. Tang et al. incorporated a frequency domain information separator and high-frequency cross-attention mechanism to ensure accurate recovery of fine details across varying dose levels, which is essential for maintaining image fidelity in clinical settings. Additionally, Han et al. [27] introduced a Contrastive Diffusion Model with Auxiliary Guidance for coarse-to-fine PET reconstruction, combining contrastive learning with diffusion-based denoising to better align low-dose PET images with high-quality targets. This approach significantly improved image correspondence, addressing limitations of tra-

ditional diffusion models in clinical PET applications.

A key drawback of the denoising process in diffusion models is its high computational cost, as it requires performing hundreds to thousands of iterative forward passes to reconstruct a single image. This makes diffusion models significantly slower compared to other deep learning approaches like GANs or autoencoders, limiting their practicality for time-sensitive applications [87].

2.3.5 Other methods

In PET imaging, the need to balance image quality with radiation exposure has led to a variety of reconstruction techniques. While CNNs, GANs, Transformer models, and diffusion models have shown significant promise in reconstructing SPET from LPET data, several other approaches have also been explored to enhance PET imaging. Semi-supervised learning is effective in scenarios where paired LPET-SPET data is limited. Jiang et al. [36] proposed a semi-supervised framework for generating SPET images using Region-Adaptive Normalization (RN) and Structural Consistency Constraint. The RN module performs region-specific normalization to handle varying intensity distributions across different regions, thus preserving anatomical boundaries. Meanwhile, the Structural Consistency Constraint ensures that structural details remain consistent throughout the generation process, effectively enhancing image quality even when using unpaired datasets. The S3PET framework [14] introduced by Cui et al. uses a two-stage approach: an unsupervised pre-training phase with dose-specific masked autoencoders (DsMAEs) and a supervised dose-aware reconstruction phase. The model incorporates modules to disentangle dose-specific knowledge, facilitating effective LPET-to-SPET reconstruction with limited paired data.

The integration of text-guided methods in PET reconstruction has emerged as an innovative approach to improve image quality by utilizing clinical information alongside imaging data. The Multi-modal Conditioned Adversarial Diffusion Model (MCAD) [17] introduced by Cui et al. leverages LPET images and clinical tabular data, such as age and weight, to guide the reconstruction of SPET images. The model em-

ploys a Multi-modal Conditional Encoder (Mc-Encoder) with Optimal Multi-modal Transport co-Attention (OMTA) to align imaging and text features, ensuring a comprehensive fusion of multi-modal data.

MRI-guided methods enhance PET reconstruction by providing complementary anatomical information that improves structural accuracy. Xiang et al. [83] used a deep auto-context convolutional neural network to combine LPET and T1-weighted MRI for estimating high-quality SPET images, refining the prediction iteratively through multiple CNN modules. STFNet [93], proposed by Zhang et al., uses a spatial-adaptive and transformer fusion network to combine MRI and PET, leveraging deformable convolutions and a transformer fusion encoder to improve the integration of spatial and global information. The method outperforms conventional U-Net approaches by maintaining texture and edge details in the reconstructed SPET images. Wang et al. [79] proposed a 3D auto-context-based locality adaptive multi-modality GAN (LA-GAN) that integrates LPET and T1-MRI using a locality-adaptive fusion mechanism, which adjusts the modality contributions dynamically at different image locations to improve PET synthesis quality.

In a more recent development, MEaTransGAN [78] combines CNN-based encoders and Transformer-based encoders to capture local spatial and global semantic features from both LPET and T1-MRI inputs, introducing an edge-aware loss to preserve fine anatomical details in the reconstructed SPET images. Additionally, PMC2-GAN [13] employs a point-based representation to better retain intricate structures, integrating MRI as an auxiliary modality to complement PET information in a flexible manner.

CT-guided techniques utilize anatomical priors from CT scans to improve PET reconstruction. Jiang et al. [37] introduced PET-Diffusion, an unsupervised latent diffusion model with CT-guided cross-attention, to better recover structural details in SPET images from LPET data.

Chapter 3

A Reverse Method of Data Augmentation for High Quality PET Image Synthesis

3.1 Introduction

PET imaging is highly valued in oncology for its significant contributions to cancer diagnosis and management [48]. However, SPET imaging involves cumulative radiation exposure, raising concerns about potential health risks [76]. Radiation doses from PET scanning are higher for the gonads, uterus, and bladder compared to other organs [32], and it is well established that radiation poses a risk to human health [53, 58]. Based on the "As Low As Reasonably Achievable" (ALARA) principle introduced by the International Commission on Radiologic Protection (ICRP) in 1977 [53], reducing the injected dosage is a promising solution to mitigate the risks associated with radiation exposure. However, LPET images, in comparison to those obtained under standard protocols, suffer from lower signal-to-noise ratio (SNR) and potentially lose anatomical details (shown in Figure 1.2). In response to these challenges, adopting computational methods for SPET synthesis from LPET has emerged as a promising direction. Here, the synthesis means generating the SPET images from the

LPET images.

Deep learning based methods have achieved great success in medical image analysis related tasks such as in segmentation, classification and, registration [44]. Methods based on deep learning have also been introduced for PET image synthesis tasks. Earlier studies have focused on using CNN [7, 69, 83] to synthesis the SPET images. Recently, GAN [24] based methods have achieved state of the art performance for PET synthesis. For example, Wang et al. [79] proposed a 3D c-GAN to generate high-quality SPET images, employing a 3D U-net-like architecture to maintain consistent information between different image types. Zhou et al. [96] and Zhao et al. [95] used a supervised CycleGAN-based architecture with a combined loss consisting of a cycle-consistency loss, Wasserstein distance loss, and a supervised learning loss. Luo et al. [46] presented the AR-GAN model, which uses an adaptive rectification network (AR-Net) and a spectral regularization term to address discrepancies and high-frequency distortions in synthesized PET images. Xue et al. [85] introduced the SS-AEGAN model, which employs an adaptive residual estimation mechanism to dynamically correct synthesized PET images and includes a self-supervised pre-training strategy to enhance feature representation. Methods based on using transformer [15, 35, 45, 75] and diffusion [22, 27, 57] have also been introduced for synthesizing PET images.

Despite their state-of-the-art (SOTA) performance, inherent to deep learning methods, particularly for medical image analysis, these methods consistently face challenges of insufficient annotated training datasets [18]. This is attributed to the expensive data acquisition and annotation procedures and ethics constraints. To mitigate these challenges, data augmentation methods that attempt to augment existing training datasets have been widely employed. Traditional augmentation methods include simple geometric transformations such as rotations, translations, and flips. However, these traditional methods usually have difficulties to introduce new imaging characteristics that are essential for feature learning [10]. More sophisticated methods involve the use of GANs to generate synthetic images. It can generate highly realistic and diverse synthetic data, surpassing traditional augmentation methods that rely on simple transformations, which enhances model robustness and mitigates class imbalances

[3]. Chaitanya et al. [8] proposed a semi-supervised task-driven data augmentation method to improve medical image segmentation. This method combined a task-driven approach with semi-supervised learning to enhance segmentation performance. In an alternative appraoch, combining data from different imaging modalities, such as PET and CT, has been shown to be able to leverage complementary information to enhance model performance. MultiRoiMix [77] introduced a method to leverage both anatomical and metabolic information from multiple modalities to create more informative training samples, which has been validated in multimodal segmentation tasks, demonstrating its effectiveness in improving model accuracy with limited data. CutMix [89] and CarveMix [94] have been developed for image classification and brain tissue injury segmentation. CutMix [89] involves cutting and pasting patches among training images to create new variations to enhance dataset diversity. Similar to CutMix [89], CarveMix [94] combines two existing labeled images to generate new labeled samples. However, CarveMix [94] distinguishes itself by being lesion-aware, focusing specifically on the lesions during the combination process and creating annotations for the generated images. These methods combine existing annotated data in innovative ways to create new training samples, enhancing the diversity and utility of the dataset. However, these methods are primarily used to build more labels, whereas the synthesis task requires more input data (LPET images). Therefore, they are not directly applicable to PET synthesis, which necessitates a different approach.

There has been limited work on data augmentation for PET synthesis. Although traditional data augmentation methods help improve model performance, they generate variations from the same original training data, limiting their ability to introduce truly novel information or improve generalizability across different data distributions. In order to enhance the model's generalization and robustness, it is crucial to incorporate additional new LPET and SPET image pairs from different patients.

To address this challenge, this research introduces a new reverse PET augmentation method to generate LPET images from SPET images. This chapter innovatively leverage the abundant SPET images that are acquired from routine PET imaging. By combining these estimated LPET images with actual LPET images, the training

dataset can be expanded with LPET images from new patients. This approach increases both the size and diversity of the LPET dataset, enhancing data augmentation and improving the quality of PET image synthesis. Moreover, it can be integrated with existing models to optimize data augmentation processes. By generating estimated LPET images, this reverse PET augmentation method effectively tackles the critical issue of data scarcity in medical imaging, contributing to the development of more robust deep learning models.

3.2 Method

The objective of this research is to address the data limitation challenge in PET image synthesis by introducing a method to generate LPET images from SPET images. This method allows us to augment the training dataset with synthetic LPET images, thus enhancing the data available for training deep learning models. The workflow shown as Figure 3.1.

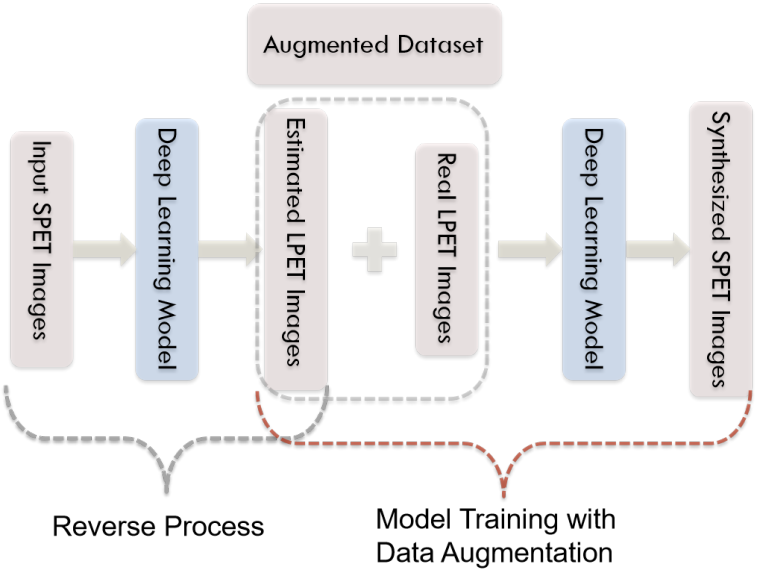


Figure 3.1 – Workflow of the Reverse Method

3.2.1 Reverse Generation of LPET Images

Given a SPET image I_S defined within a spatial domain $\Omega \subseteq \mathbb{R}^3$, this method generates an estimated low-dose PET image I'_L. The process will produce I'_L that closely mimics the characteristics of true low-dose images. This process can be mathematically represented as:

$$I'_L = G_{HL}(I_S) \tag{3.1}$$

where G_{HL} is the model trained to generate low-dose images from standard-dose inputs.

3.2.2 Integration with Real LPET Data

To enhance the training dataset, I combine these estimated LPET images I'_L with real LPET images I_L. This integration increases the diversity and volume of the training data, which is crucial for effective data augmentation. The combined dataset D can be expressed as:

$$D = \{(I'_L, I_S), (I_L, I_S)\} \tag{3.2}$$

where (I'_L, I_S) pairs the generated low-dose images with their corresponding standard-dose images, and (I_L, I_S) pairs the real low-dose images with their standard-dose counterparts.

3.2.3 Training with Augmented Data

The augmented dataset D is then used to train a deep learning model for PET image synthesis. The model architecture of both from LPET to SPET and from SPET to LPET is the same. The training process involves minimizing a loss function that

measures the discrepancy between the predicted standard-dose images I'_S and the actual standard-dose images I_S. This process can be mathematically represented as:

$$I'_S = G_{LH}(D) \tag{3.3}$$

where G_{LH} is the model trained to generate standard-dose images from low-dose inputs.

3.2.4 Loss Function and Optimization

I employ the L1 Loss function, which is defined as:

$$\mathcal{L}_{HL} = \frac{1}{N}\sum_{i=1}^{N}|I_L^{(i)} - I_L'^{(i)}| \tag{3.4}$$

$$\mathcal{L}_{LH} = \frac{1}{N}\sum_{i=1}^{N}|I_S^{(i)} - I_S'^{(i)}| \tag{3.5}$$

where N is the total number of pixels in each image, $I_S^{(i)}$ and $I_S'^{(i)}$ represents the pixel values of the target and estimated SPET images, $I_L^{(i)}$ and $I_L'^{(i)}$ represents the pixel values of the target and estimated LPET images. The equation Equation 3.4 is the loss function for the process G_{HL} and equation Equation 3.5 is the loss function for the process G_{LH}. The optimization of the model parameters is performed using the Adam optimizer, which adjusts the model parameters to minimize the L1 loss:

$$\theta_{HL} = \arg\min_{\theta_{HL}} \mathcal{L}_{L1}(I_L, I'_L) \tag{3.6}$$

$$\theta_{LH} = \arg\min_{\theta_{LH}} \mathcal{L}_{L1}(I_S, I'_S) \tag{3.7}$$

where θ_{HL} and θ_{LH} denote the model parameters.

3.2.5 Integration with State-of-the-Art Models

This proposed method can be seamlessly integrated with any SOTA model for PET image synthesis. By incorporating the generated LPET images into the training process, I enhance the model's ability to generalize and improve its performance in synthesizing high-quality PET images.

3.3 Experiment Setup

3.3.1 Dataset

Ultra-low Dose PET Imaging Challenge dataset was used for experiments [1]. A total of 387 patient studies were acquired using the Siemens Biograph Vision Quadra with Time-of-Flight (TOF). Images were reconstructed using Ordered Subset Expectation Maximization (OSEM) at a resolution of $440 \times 440 \times 644$ pixels, with a voxel spacing of $1.65\,\mathrm{mm}^3$. All acquired data was in 'list mode,' enabling reconfiguration of acquisition duration to simulate various low-dose scenarios[1]. Each LPET scenario was characterized by a dose reduction factor (DRF) derived from reconstructed counts over a reduced acquisition time frame centred around the midpoint of the original acquisition duration. Low-dose images were generated using DRFs of 2, 4, 10, 20, 50, and 100, in conjunction with a corresponding standard-dose image, to cover a comprehensive range of dose levels. In this study, we mainly focused on using a DRF of 100 as our training input.

3.3.2 Implementation Details

All experiments were conducted using PyTorch with TensorBoard for visual analytics, using an i7-5930K CPU and a 24 GB NVIDIA RTX A3090 GPU. Optimization was carried out using the ADAM optimizer[41]. In our training, the learning rate was set to 0.0002, batch size was set to 2 and we trained our model for over 84,000 iterations.

3.3.3 Experimental Settings

We randomly selected 40 patients to simulate a small clinical dataset for training purposes, where we trained the models M_{LH} (LPET to SPET) and M_{HL} (SPET to LPET). Both models share the same architecture. Additionally, we allocated 10 patients for validation, maintaining consistency in datasets across models.

For further training based on our reverse data augmentation method, we utilized 224 additional patients, assuming we only had SPET images for these cases. The model M_{HL} was employed to estimate the corresponding LPET images. We then combined these estimated 224 LPET images with the 40 real LPET images to train the final model.

To compare the impact of dataset size, we also used all 284 patients (244+40=284) with DRF100 (low dose) data and their standard-dose data to simulate a large clinical dataset. This allowed us to evaluate the differences between training with small and large datasets and to determine if the reverse data augmentation method can improve the results.

For testing, 75 patients were allocated, and this dataset was used to evaluate all models. Also, we allocated 38 patients for validation.

3.3.4 Evaluation Metrics

To evaluate the performance of the PET image synthesis models, we utilized several standard metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Mean Squared Error (MSE), and Normalized Root Mean Squared Error (NRMSE). These metrics are widely used in image processing to quantify the quality and similarity of images.

Peak Signal-to-Noise Ratio (PSNR)

PSNR is a measure of the peak error between the synthesized image and the reference image. It is defined as:

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{MAX}_{I_{ref}}}{\sqrt{\text{MSE}}} \right) \tag{3.8}$$

where $\text{MAX}_{I_{ref}}$ is the maximum possible pixel value of the reference image. Higher PSNR values indicate better image quality.

Structural Similarity Index (SSIM)

SSIM is used to measure the similarity between two images. It considers changes in structural information, luminance, and contrast. The SSIM index is defined as:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3.9}$$

where μ_x and μ_y are the mean intensities, σ_x and σ_y are the standard deviations, and σ_{xy} is the covariance of x and y. C_1 and C_2 are constants to stabilize the division. Higher SSIM values indicate better structural similarity.

Mean Squared Error (MSE)

MSE measures the average squared difference between the synthesized image and the reference image. It is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (I_{\text{syn}}(i) - I_{\text{ref}}(i))^2 \tag{3.10}$$

where I_{syn} is the synthesized image, I_{ref} is the reference image, and n is the number of pixels. Lower MSE values indicate better image quality.

Normalized Root Mean Squared Error (NRMSE)

NRMSE normalizes the root mean squared error by the range of the image intensities, providing a relative measure of the error. It is defined as:

$$\text{NRMSE} = \frac{\sqrt{\text{MSE}}}{\text{MAX}_{I_{ref}} - \text{MIN}_{I_{ref}}} \tag{3.11}$$

where MAX_I and MIN_I are the maximum and minimum possible pixel values of the reference image, respectively. Lower NRMSE values indicate better image quality.

These metrics collectively provide a comprehensive assessment of the image synthesis quality, capturing different aspects of image fidelity and similarity.

3.3.5 Comparison Method

Our reverse data augmentation method was evaluated with several state-of-the-art synthesis methods including the traditional 3D-UNet [11] and four GAN-based methods specifically designed for PET synthesis—cGAN [79], Cycle-GAN [95, 96], AR-GAN [46], and SS-AEGAN [85]. The number of features at each level for different models were as follows: the 3D-UNet has feature counts of $[16, 32, 64, 128, 256]$, the cycleGAN has $[16, 32, 64, 128]$, and the cGAN has $[16, 32, 64, 128, 256]$. Additionally, the generator filters in the first convolutional layer were set to 32 for AR-GAN and 64 for SS-AEGAN. All models were implemented using PyTorch.

3.4 Results and Discussion

This method involved training models with both low-to-high and high-to-low dose data transformations. The results, summarized in Table 3.1, and the visualization in Figure 3.2, demonstrate that the proposed reverse data augmentation method enhanced the performance of all these models for PET image synthesis. The figure shows that for AR-GAN, the visualization on the brain using reverse data appears

clearer than using only the small dataset. For SS-AEGAN, the boundaries are more distinct and less blurry compared to the visualization generated from only the small dataset. The table compares all the models using the small dataset, large dataset and reversed dataset.

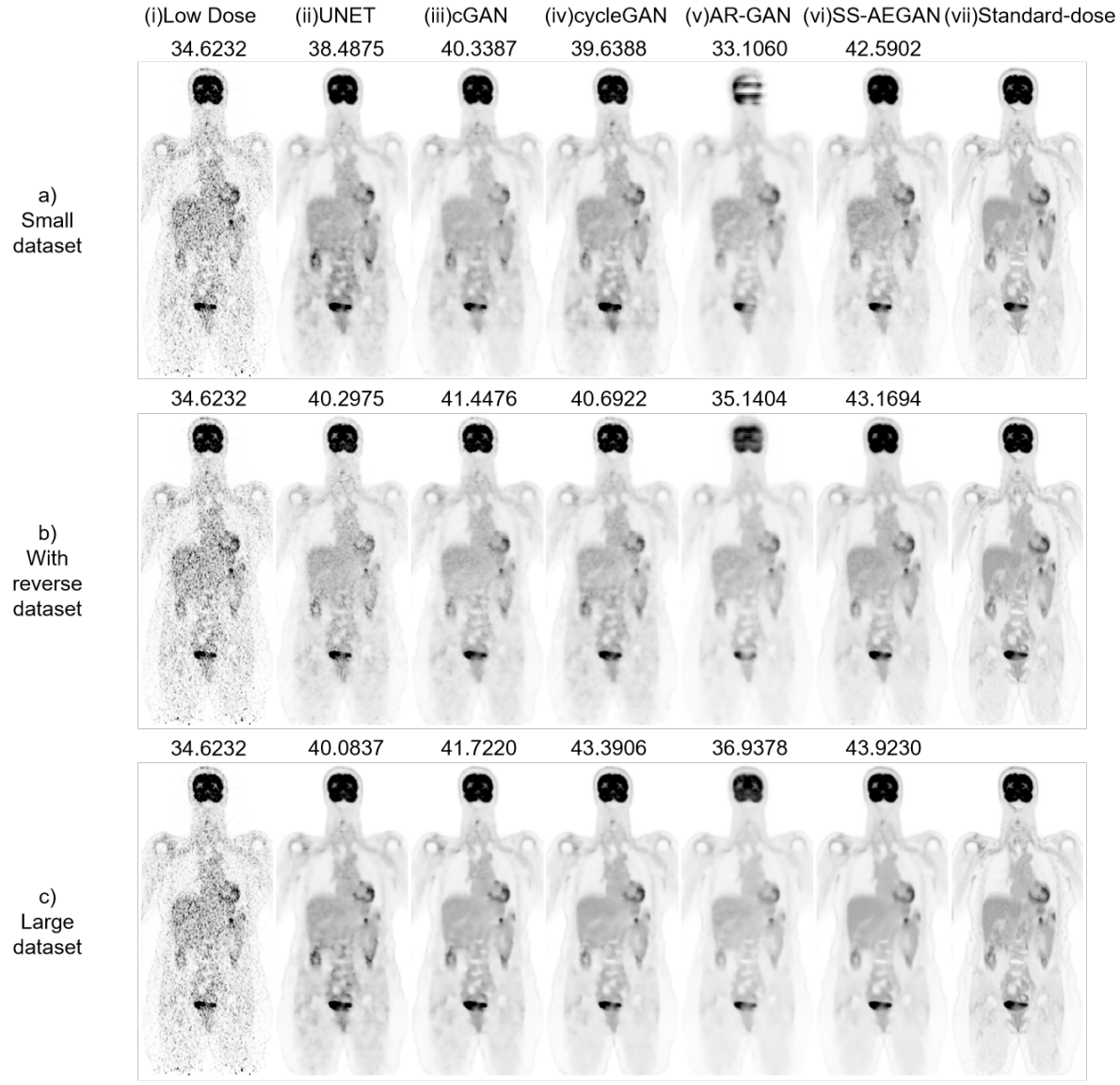


Figure 3.2 – Visualization of an example patient: The value in the figure is the PSNR value(dB) for the example patient.

Table 3.1 – Comparison of different models for Reverse Strategy

		PSNR(↑)(dB)	SSIM(↑)	MSE(↓)	NRMSE(↓)(%)
RAW_DRF100		42.6927 ± 6.1795	0.9141 ± 0.1003	0.3235 ± 0.3038	0.8992 ± 0.6784
UNET	small dataset	47.4783 ± 4.5450	0.9834 ± 0.0149	0.1229 ± 0.1722	0.4854 ± 0.2673
	large dataset	**49.2332 ± 5.0233**	**0.9876 ± 0.0147**	**0.0646 ± 0.0416**	**0.4112 ± 0.2649**
	with reverse data	47.9772 ± 5.0870	0.9810 ± 0.0196	0.0947 ± 0.0957	0.4740 ± 0.2937
cGAN	small dataset	47.1607 ± 4.3898	0.9822 ± 0.0160	0.1287 ± 0.1674	0.4985 ± 0.2630
	large dataset	**49.9077 ± 4.9942**	**0.9899 ± 0.0111**	**0.0582 ± 0.0449**	**0.3792 ± 0.2420**
	with reverse data	49.1327 ± 4.8766	0.9863 ± 0.0144	0.0718 ± 0.0650	0.4112 ± 0.2547
cycleGAN	small dataset	47.9932 ± 4.6849	0.9837 ± 0.0139	0.1117 ± 0.1672	0.4626 ± 0.0027
	large dataset	**50.5180 ± 4.8456**	**0.9908 ± 0.0124**	**0.0483 ± 0.0299**	**0.3511 ± 0.2253**
	with reverse data	48.5172 ± 4.6989	0.9871 ± 0.0127	0.0790 ± 0.0556	0.4356 ± 0.2565
AR-GAN	small dataset	43.7500 ± 5.2979	0.9839 ± 0.0120	0.3306 ± 0.4329	0.7818 ± 0.5022
	large dataset	**47.3284 ± 4.7104**	**0.9896 ± 0.0156**	**0.0950 ± 0.0474**	**0.5000 ± 0.2951**
	with reverse data	45.4202 ± 5.1001	0.9814 ± 0.0167	0.1675 ± 0.1506	0.6350 ± 0.3823
SS-AEGAN	small dataset	48.3442 ± 3.7591	0.9784 ± 0.0142	0.1027 ± 0.1204	0.4211 ± 0.1979
	large dataset	**51.8235 ± 4.7664**	**0.9902 ± 0.0152**	**0.0360 ± 0.0225**	**0.3013 ± 0.1962**
	with reverse data	50.0932 ± 3.9961	0.9829 ± 0.0178	0.0599 ± 0.0501	0.3514 ± 0.1954

[a] 'small dataset' is the model trained with data from 40 patients, converting LPET images to SPET images.
[b] 'large dataset' is the model trained with data from all 284 patients, converting LPET images to SPET images.
[c] 'with reverse data' is the model trained trained with 40 real LPET and 244 estimated LPET images.
[d] The cell in bold represents the best result, while the underlined cell indicates the second best.

3.4.1 Analysis about the large dataset and the small dataset.

By comparing the small dataset to the large dataset, as expected, significant improvements were observed across all tested methods. For instance, even the best-performing model, SS-AEGAN, demonstrated notable enhancements: PSNR increased from 48.3442 dB to 51.8235 dB, SSIM rose from 0.9784 to 0.9902, MSE dropped from 0.1027 to 0.0360, and NRMSE decreased from 0.4211% to 0.3013%. These improvements underscore the importance of training data volume, as larger datasets provide more diverse examples and richer patterns for the models to learn. This finding aligns with the conclusion of Sun et al. [70], which emphasized that training with larger datasets leads to more effective models and higher-quality synthesized SPET images. Furthermore, the performance gap highlights the limitations of models trained on small datasets, which are more prone to overfitting and may fail to generalize well across unseen data.

3.4.2 Analysis about the reverse data augmentation method compared to the small dataset

With the additional data created from this reverse data augmentation method, this inclusion consistently enhanced the performance of all the testing models. For instance, the SS-AEGAN model with reverse data augmentation achieved the highest PSNR of 50.0932, indicating a large improvement in image quality. This model also exhibited the lowest MSE and NRMSE values, confirming its superior accuracy and reduced error margins. Similarly, the UNET and cGAN models demonstrated noticeable enhancements in PSNR and SSIM, underscoring the effectiveness of this augmentation method across different architectures.

The visualization in Figure 3.2 further corroborates these quantitative findings. The synthesized PET images augmented with reverse data exhibit clearer anatomical details and richer texture information, especially for the AR-GAN method. The SS-AEGAN model, in particular, demonstrated enhancements in SSIM values, indicating

that the structural similarity of the synthesized images is markedly improved by the reverse data augmentation method.

Moreover, the comparative analysis demonstrated that the models trained with the proposed reverse data augmentation method consistently outperformed those trained with small datasets alone, regardless of their initial performance levels. For example, the SS-AEGAN model, which already performed the best among the models without reverse data, still showed improvements when trained with the reverse data augmentation method. Specifically, the PSNR increased from 48.3442 dB to 50.0932 dB, SSIM improved from 0.9784 to 0.9829, MSE decreased from 0.1028 to 0.0588, and NRMSE reduced from 0.4211% to 0.3514%. These enhancements across all metrics indicate that the reverse data augmentation method can boosts the models' generalizability and robustness.

3.4.3 Analysis of the reverse data augmentation method compared to the large dataset.

Compared with the results of models trained on large dataset that includes a comprehensive low-dose and standard dose pair dataset, the performance of using the large dataset was noticeable. From Figure 3.2, AR-GAN failed in brain synthesis, producing images with noticeably blurred and distorted brain regions. The anatomical structure lacks clarity, and key features are either smoothed out or missing, indicating poor synthesis quality. This suggests that AR-GAN struggles to capture the fine-grained details and complex spatial relationships required for accurate brain region synthesis on limited data. Nonetheless, for the cGAN model, the performance with reverse data augmentation method was very close to that with all the low-dose data, highlighting the potential of the reverse data augmentation method to enhance the quality of synthesized SPET images. However, cycleGAN exhibited the largest discrepancy between the large dataset and the reverse dataset, as shown in Table 3.1, indicating that cycleGAN is highly influenced by dataset size. Despite this, the reverse data augmentation method has its limitations, as evidenced by the noticeable

performance gap between the large dataset and the reverse dataset. While small dataset integrate the reverse dataset outperforms the small dataset alone, it still falls short of the results achieved by directly training on the large dataset, indicating that reverse data cannot fully substitute for real, diverse training samples. While cycleGAN can improve the synthesis of SPET images, it does not necessarily enhance the precision of estimated LPET images. Generating precise LPET images is challenging, and the estimated LPET images may contain artifacts; for example, the generated images often appear over-smoothed. This contributes to the substantial difference observed between the large dataset and the reverse dataset for cycleGAN.

3.4.4 Overall Discussion

These findings underscore the potential of reverse data augmentation method in addressing the limitations posed by small datasets in medical imaging. The proposed reverse data augmentation method offers a practical solution to improve outcomes with limited data. By augmenting the training data with reverse transformations, this method increases the diversity and variability of the dataset, which is crucial for developing reliable and effective deep learning models. This is particularly important in PET imaging, where data diversity directly impacts the accuracy and robustness of diagnostic tools. Notably, the improvements in MSE and NRMSE metrics indicate that not only the mean values but also the variance of these metrics have been improved, suggesting enhanced robustness of the results.

3.5 Summary

In this study, a reverse data augmentation method was introduced to augment the training dataset by levering the SPET images. The experimental results with a clinical dataset demonstrate that the reverse data augmentation method enhances the performance of PET synthesis models when combined with state-of-the-art techniques. By leveraging both low-to-high and high-to-low dose data transformations, this method

effectively mitigates the limitations of small datasets. Although a gap remains when compared to using a comprehensive dataset containing all low-dose data, the reverse data augmentation method still substantially improves the synthesized image quality with limited datasets. Future work will focus on developing more precise methods for generating LPET images to further improve the data augmentation process and exploring the applicability of this approach to other medical imaging tasks.

Chapter 4

TCCA-Net: A Time-Controlled Channel-Spatial Attention Enhanced ResUNet for PET image synthesizes

4.1 Introduction

PET imaging, detailed earlier in this thesis, utilizes radioactive tracers to visualize metabolic processes and physiological activities across oncology, cardiology, and neurology applications [4, 48]. Despite its clinical value, concerns about radiation risks from repeated PET scans [76], particularly in oncology patients, necessitate reducing injected tracer doses, leading to LPET. However, LPET images typically suffer from lower SNR. A promising solution, as previously introduced, is the synthesis of high-quality SPET images from LPET images using advanced image processing techniques.

In recent years, substantial progress has been made in addressing the challenge of synthesis high quality PET image from low-dose PET. Building on this progress, deep

learning models have played a critical role in advancing image synthesis techniques, particularly for enhancing LPET images. CNN based methods have become foundational in image synthesis techniques. The U-Net [59] and 3D U-Net [11], initially developed for biomedical image segmentation, have also been effectively applied to PET image synthesizing [64]. Building on the concept of residual learning [92], many researchers built models with the residual learning [21, 65, 84]. Sano et al. [63] used 2D residual U-Net to do the PET image denoising. Spuhler et al. [69] proposed a 2D residual dilated CNN to predict the SPET images from LPET counterpart. GAN [24] have become a popular baseline in image synthesis due to their ability to learn effectively from paired image data. GANs consist of a generator, which creates realistic synthetic images, and a discriminator, which differentiates between real and synthetic images through adversarial training, leading to continuous improvement of image realism. Yan Wang et al. [79] implemented a 3D c-GAN with a U-Net-like generator for SPET synthesis. Zhou et al. [96] and Zhao et al. [95] used CycleGAN architectures, adding supervised loss functions to enhance LPET recovery. Luo et al. [46] introduced AR-GAN, which incorporated adaptive rectification to refine SPET synthesis outputs and spectral regularization to preserve high-frequency details, ensuring alignment in the frequency domain. Xue et al. [86] proposed CG-3DSRGAN which incorporates a classification module to enhance the accuracy of low-to-high-dose translation by utilizing a multi-task generator and a secondary network for recovering spatial details.

However, these methods expand the receptive field through down-sampling, but struggle to capture long-range correlations [68, 97]. To address this challenge, transformer-based methods have been applied. TriDo-Former [15], TriDoRNet [38], and PK-TriDo [16] integrate various domain information to enhance the quality of synthesized images. Transformers [75] benifits for their self-attention mechanisms for capturing long-range dependencies. In order to capture both global and local features for generating high-quality SPET images, Luo et al. [45] and Zeng et al. [91] proposed models that combine Transformer with CNN or GAN methods. However, these spatial-wise attention in Transformer models is computationally and memory-intensive, limiting

their application to lower image resolutions and restricting detailed textural representation at full resolution. To address this, researchers have introduced channel-wise self-attention methods [2, 35, 90] to balance efficiency and feature detail, with Jang et al. [35] further combining spatial- and channel-wise attention. However, transformer architectures rely heavily on global self-attention mechanisms. These mechanisms typically require large amounts of data to learn meaningful relationships, making Transformers less effective at modeling detailed local spatial contexts and subtle textures present in limited datasets [66], leading to convergence challenges and high memory consumption, necessitating further optimization to enhance their efficiency across diverse applications. Moreover, most existing methods are designed for PET images acquired at a specific dose level; for instance, models trained on DRF20 images are typically tested only at the same DRF20 level. This approach restricts their ability to adapt effectively to the diverse dose conditions commonly encountered in clinical practice. This rigid design restricts their real-world applicability, as models trained on one dose level may fail to generalize across others. Addressing this limitation is a key focus of this study.

In existing literature, models for PET image synthesis are typically trained on datasets acquired at fixed dose levels. However, in clinical practice, patient-specific factors—such as body weight, metabolic rate, and overall health—necessitate individualized PET imaging protocols, leading to variable dose administrations [60]. This variability can result in dose levels that differ from those represented in standard training datasets, potentially affecting the performance of models trained under fixed-dose assumptions. For instance, a model trained exclusively on images obtained at a 1% dose level (DRF100) may not perform optimally when applied to images acquired at different dose levels, such as 2%, 5%, or 10%. This discrepancy underscores the importance of developing models capable of adapting to a range of dose conditions, thereby enhancing their robustness and clinical applicability. So the model must generalize across different noise levels rather than being overfitted to a specific dose level. Thus, an algorithm is required that can denoise images across arbitrary noise levels rather than just one. To tackle cross-tracer and cross-scanner variability, Xue et al.

[85] proposed an adaptive multistage GAN model, which achieved better performance on LPET images at different dose levels. However, the model needed to input different dose-level PET images for training, limiting its practicality in scenarios where such diverse dose-level datasets are unavailable.

In this study, a TCCA-Net is proposed - A Time-Controlled Channel-Spatial Attention Enhanced ResUNet, to help with the PET image synthesizing across varying dose levels by addressing key challenges in biomedical image generation. This chapter introduce a Time-Controlled Mechanism, a novel approach to enhance PET image synthesis by explicitly modeling noise variations across different dose levels. Unlike conventional methods that train on fixed-dose PET images, this approach simulates PET images at multiple noise levels by introducing controlled noise into target images, with the noise level regulated by a timestep t. The model then computes the loss against LPET images, using t as an additional input to dynamically guide SPET synthesis. Before introducing the time-controlled mechanism, Full-TrSUN, a full-resolution hierarchical transformer framework, is introduced. This approach innovatively employs hierarchical transformer blocks to distill multi-scale features directly from the full image resolution. The process integrates XCIT Transformer Blocks [2] with a CNN encoder-decoder architecture and UNet connections [59], collectively enabling the model to capture a comprehensive range of textural nuances. The use of the CNN structure also captures spatial relationships and enhances efficiency, resulting in higher performance. Then, building upon the previously introduced Full-TrSUN, the goal is to enhance feature representation by effectively capturing both local and global features within an image. So, an Attention-Enhanced ResUNet (AE-ResUNet), which integrate both channel-wise and spatial-wise attention mechanisms to enhance feature extraction, was introduces. To enhance gradient flow, feature preservation, and learning stability, the ResNet block was integrated alongside spatial and channel attention. Given that low-to-high datasets are typically much smaller than natural image datasets, this hybrid design also helps improve convergence by balancing local feature extraction and global context learning.

This timestep-aware framework allows the model to learn a continuous representation

of noise across varying dose conditions, rather than being restricted to discrete training datasets. By incorporating t, the model can better adapt to unseen dose levels, effectively bridging the gap between different noise conditions. This strategy not only enhances robustness but also introduces a more flexible and scalable approach to PET image denoising—offering a significant advancement over existing fixed-dose training paradigms.

This chapter conducted a series of experiments to evaluate the effectiveness of this proposed approach using Ultra-low Dose PET Imaging Challenge [1]. This results demonstrate significant improvements over existing methods in terms of PSNR, MSE, NRMSE and SSIM.

4.2 Method

The objective of this study is to generate standard-dose PET images I_S from low-dose PET images I_L. Both I_L and I_S are defined as three-dimensional (3D) volumes situated within a spatial domain $\Omega \subseteq \mathbb{R}^3$. The proposed model consists of three key components.

First, to capture features from full image resolution, Full-Trsun[72] is used - a Full-Resolution Transformer UNet. This model integrates a transformer encoder alongside a CNN encoder, followed by a CNN decoder that leverages the hidden state and UNet connections. The architecture of Full-Trsun is illustrated in Figure 4.1

Second, to enhance feature representation, this study proposed an AE-ResUNet, which incorporates both channel-wise and spatial-wise attention mechanisms. Furthermore, ResNet blocks are integrated to improve gradient flow, feature preservation, and learning stability. The overall pipeline is depicted in Figure 4.2.

Finally, to enhance the generalizability of the model, a time-controlled mechanism was introduced. This mechanism utilizes a timestep to represent image noise levels, allowing the model to be evaluated across different noise conditions, ultimately improving its robustness. The final model structure of TCCA-Net is shown in Figure 4.4.

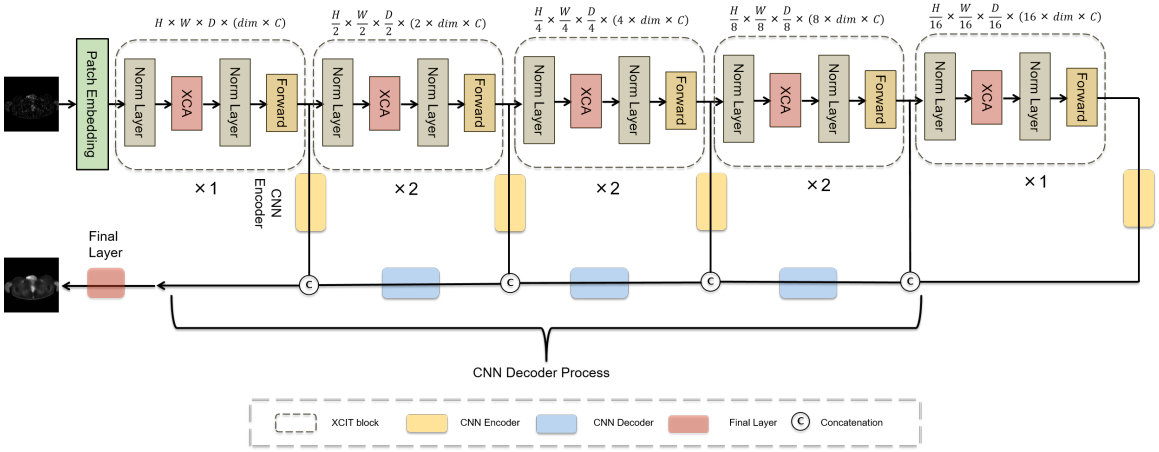


Figure 4.1 – Architecture of the Full-TrSUN model.

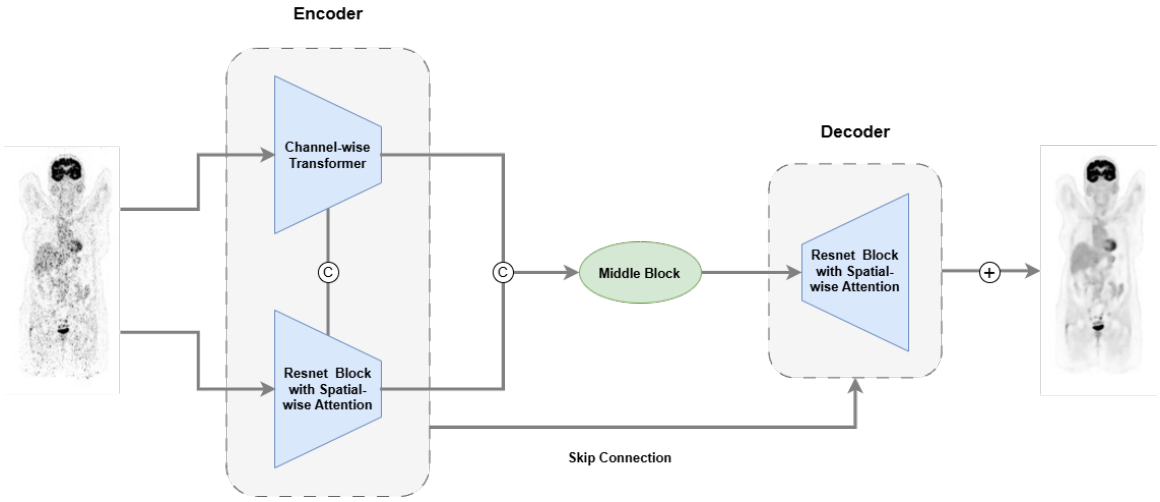


Figure 4.2 – Pipeline of AE-ResUNet.

4.2.1 Full-TrSUN

Initially, the process utilizes Transformer encoder as a foundational hidden block. Assuming the presence of t distinct stages within this block, each stage produces a downsampled output. At the i^{th} level $i \in \{1, 2, \cdots, t\}$, the resultant output is denoted as I_L^i which employed as inputs for a CNN-encoder block, denoted as I_E^i. The output is then directed through a CNN decoder block, facilitated by skip connections at various resolutions to effectively maintain and integrate multi-scale features. Upon completion of the decoding process through the final layer, an estimated standard-dose PET image I_S' is produced.

The architecture of the Full-TrSUN is illustrated in Figure 4.1, which consists of a transformer encoder stage combined with the CNN encoder process followed by the CNN decoder using the hidden state and UNet connection.

Transformer Stage

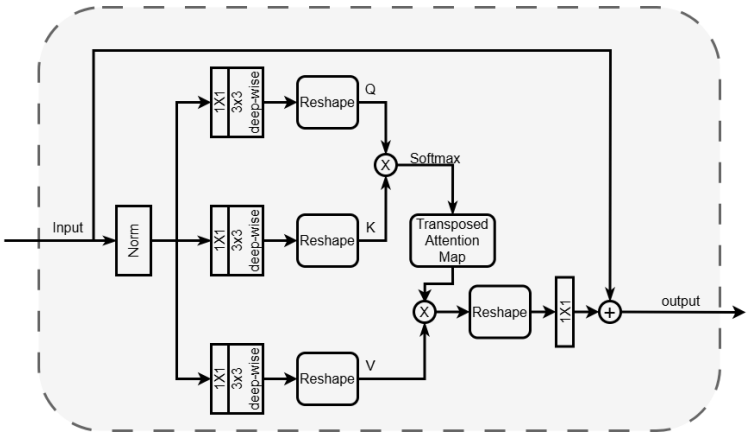


Figure 4.3 – channel-wise attention structure.

Starting with the low-dose PET image $I_L \in \mathbb{R}^{H \times W \times D \times C}$ as the input, the transformer stage initially employs a $3 \times 3 \times 3$ convolution layer to obtain the low-level feature embeddings. Following the patch embedding, the input image retains its full resolution,

ensuring the extraction of complete textural information at the tissue level. The entire encoder process comprises five Transformer Stages, each with a different number of transformer blocks arranged as (1,2,2,2,1). A critical precursor in this methodology is the initial convolutional transformation of an LPET input image $I_L \in \mathbb{R}^{H\times W\times D\times C}$ to extract foundational feature representations $F_0 \in \mathbb{R}^{H\times W\times D\times (dim\times C)}$, where H, W, and D delineates the spatial dimensions, and $dim\times C$ denotes the channel count. This extraction lays the groundwork for subsequent transformations across five distinct levels, thereby evolving F_0 into enriched feature embeddings $F_d \in \mathbb{R}^{H\times W\times D\times (d\times dim\times C)}$.

Two components of the Transformer block are Cross-Covariance Attention (XCA) and Forward process[2]. The XCA mechanism is articulated as,

$$\hat{\mathbf{X}} = WAttention(\mathbf{Q},\mathbf{K},\mathbf{V}) + \mathbf{X} \tag{4.1}$$

$$Attention(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{V}Attention(\mathbf{Q},\mathbf{K}) = \mathbf{V}\cdot Softmax(\frac{\mathbf{K}^T\mathbf{Q}}{\alpha}) \tag{4.2}$$

where $\mathbf{X}$ and $\hat{\mathbf{X}}$ are the input and output features map and $\mathbf{Q}(query)$, $\mathbf{K}(key)$, $\mathbf{V}(value)$ are obtained after reshaping tensors from the original size. Here $\mathbf{Q} = \mathbf{X}W_q$, $\mathbf{K} = \mathbf{X}W_k$ and $\mathbf{V} = \mathbf{X}W_v$, which W_q, W_k, W_v are the weights. It shifts the computational emphasis from the spatial domain to the channel domain, thereby achieving a linear complexity. Depth-wise convolutions complement this shift to underscore local context before the computation of the global attention. Here α is the scaling factor.

After the XCA block, a Forward block was used to enhance the communication between each XCA block by two $3\times 3\times 3$ convolution layers with Batch Normalization and GELU non-linearity. Figure 4.3 shows the structure of the XCA block, which represents the channel-wise attention mechanism.

CNN Encoder and CNN Decoder

This model extracts a sequence of representations F_i ($i \in \{1,2,\ldots,5\}$), each with dimensions $\frac{H}{P}\times\frac{W}{P}\times\frac{D}{P}\times C$, where $P = 2^i, C = dim\times i$. This study choose the fea-

ture size and the special dim with 24 and 3 for the CNN-Encoder and CNN-Decoder blocks. For the Encoder block, this study use these representations as the input of the encoder blocks with two $3 \times 3 \times 3$ convolutional layers, each followed by a normalization layer. For the Decoder block, the spatial resolution of these feature maps is subsequently amplified by a factor of 2 via a deconvolutional layer. Utilizing these convolutional layers enables the effective projection of complex, high-dimensional data into a more interpretable and spatially relevant form, laying the groundwork for enhanced feature synthesis and integration within the decoder through skip connections at corresponding levels. The synthesized features are further processed in another decoder block, adhering to the previously described configuration and including an upsampling phase. The final outputs are then generated by applying a $1 \times 1 \times 1$ convolutional layer, followed by a sigmoid activation function, thus enabling the effective synthesis task.

This study use the hidden state output as the output of the Transformer Block to connect the Transformer Block and the CNN-encoder block. Then, this study use the concatenation of the Encoder output as the Decoder input like UNet[59].

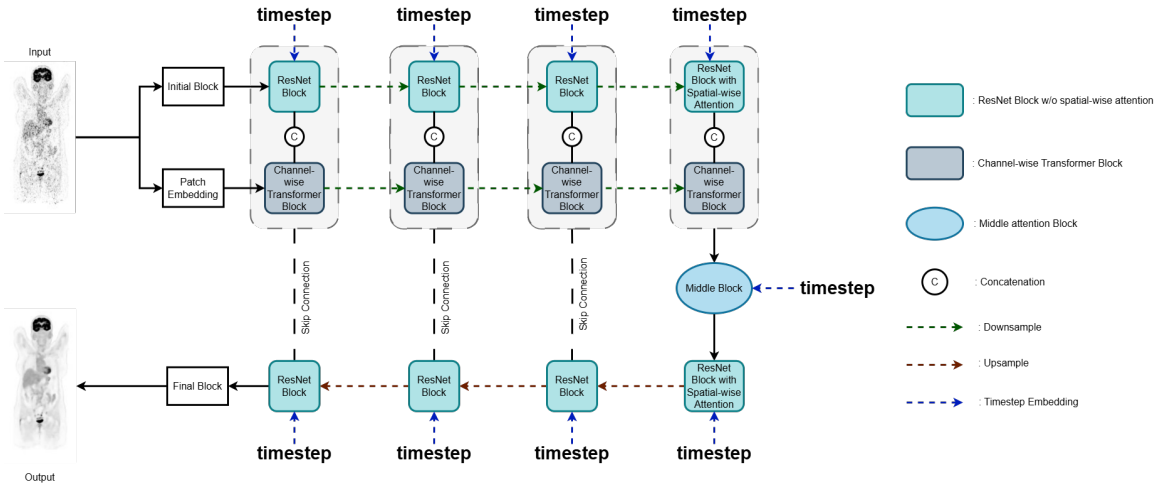


Figure 4.4 – Architecture of TCCA-Net.

4.2.2 Attention-Enhanced ResUNet(AE-ResUNet)

Building upon Full-TrSUN, we propose the AE-ResUNet, features a dual-encoder architecture. The first encoder follows Full-TrSUN, incorporating channel-wise attention to enhance feature selection across channels. The second encoder is based on ResUNet, consisting of two ResNet blocks, with the final block integrating spatial attention to refine spatial feature representation. This design improves gradient flow, enhances learning stability, and strengthens both channel-wise and spatial-wise feature extraction.

Here a Dual-Encoder Structure was used. Let $I_L \in \mathbb{R}^{H\times W\times D}$ be the low-resolution input, where H, W, D represent spatial dimensions.

The first encoder follows the Full-TrSUN design, incorporating channel-wise attention to refine feature selection across different channels. For the formula related to the first decoder, refer to Section 4.2.1. At the i^{th} level $i \in \{1, 2, \cdots, n\}$, the resulting output is denoted as I^i_{CAE}, which is used as input for the next channel-wise attention encoder (CAE) block.

The second encoder follows the ResUNet design, integrating spatial attention in the final block to enhance spatial feature representation. At the i^{th} level, $i \in \{1, 2, \cdots, n\}$, the resulting output is denoted as I^i_{RAE}, which is used as input for the next ResNet-Attention Encoder (RAE) block.

$$I^i_{RAE} = I^{i-1}_{RAE} + f(I^{i-1}_{RAE}) \tag{4.3}$$

where:

- $f(I^{i-1}_{RAE})$ represents the residual function (convolution, batch normalization, ReLU),
- I^{i-1}_{RAE} is the input feature map.

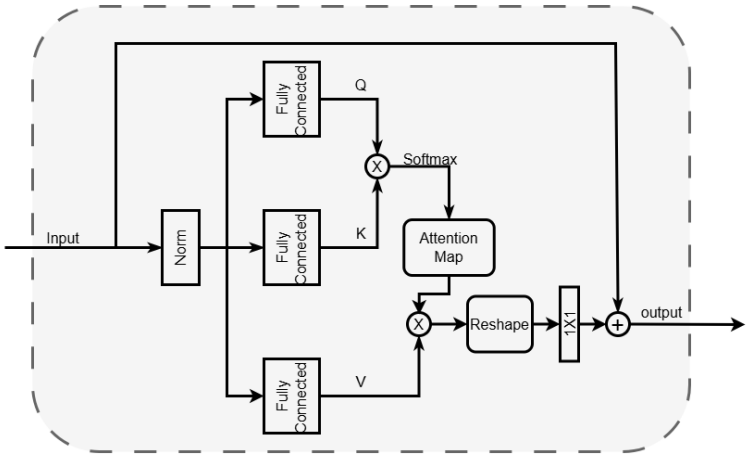


Figure 4.5 – Spatial-wise Attention structure.

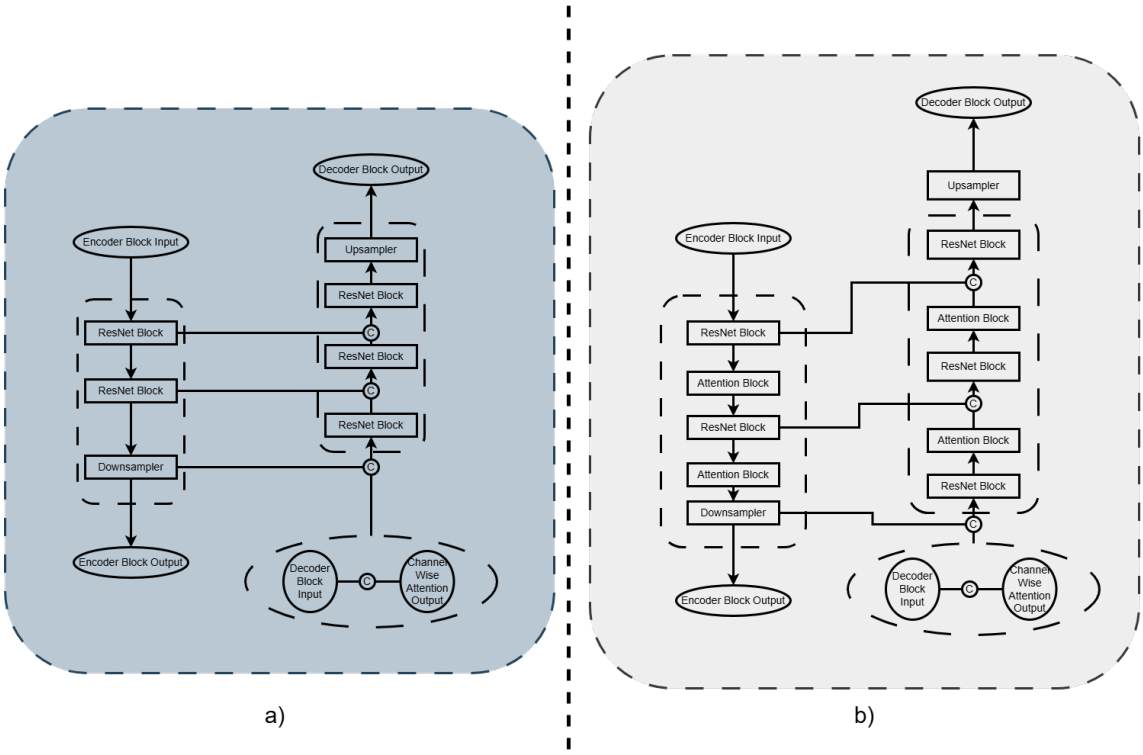


Figure 4.6 – The figure illustrates the structure of the AE-ResUNet block, where only the last layer includes spatial attention. (a) represents the block without attention, while (b) represents the block with attention.

At the final ResNet block, spatial attention is applied. The spatial-wise attention[75] is articulated as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.4}$$

where:

- $Q \in \mathbb{R}^{m \times d_k}$ is the matrix of query vectors.
- $K \in \mathbb{R}^{n \times d_k}$ is the matrix of key vectors.
- $V \in \mathbb{R}^{n \times d_v}$ is the matrix of value vectors.
- d_k is the dimensionality of the keys and queries.

The structure of the spatial-attention is shown in Figure 4.5

The final layer outputs from both encoders are concatenated and used as the input to the middle block. Similarly, the intermediate layer outputs of the encoders are concatenated with the decoder's output to serve as the input for the next decoder stage. The skip connection way is shown in Figure 4.6

4.2.3 Time-controlled mechanism

To improve the generalizability and robustness of this model in synthesizing high-quality PET images, this study introduces a time-controlled mechanism that dynamically adapts to varying noise levels in the target SPET images. This mechanism leverages a timestep variable t randomly sampled at each training iteration to control the noise added to the target image. By incorporating this dynamic noise simulation, the model learns to reconstruct high-quality PET images that align with the SPET target.

During training, noise was introduced into the high-quality SPET image I_S using a randomly sampled timestep t, generating a degraded version I_t that simulates different noise levels:

$$I_t = I_S + \epsilon_t \tag{4.5}$$

where:

- $\epsilon_t \sim \mathcal{N}(0, \sigma_t^2 I)$ is Gaussian noise controlled by t.
- σ_t is a noise-dependent scaling factor.
- I_t represents the artificially degraded PET image, used to measure the discrepancy with the LPET image.

The gap between I_t and I_L(LPET images) was estimated through a **loss function**, guiding the model in learning the optimal transformation from LPET to high-quality SPET.

The input to the model consists of:

- The **LPET image** I_L, which the model transforms into a high-quality PET image.
- The **timestep** t, which encodes the noise level added to I_S and is used within the model to guide feature extraction.

To effectively incorporate the timestep t into the network, I use a sinusoidal embedding function inspired by Ho et al. [31]. This embedding transforms the discrete timestep into a continuous representation that the model can use to condition the feature extraction process dynamically.

The timestep embedding function is defined as:

$$\tau(t) = \text{Embedding}(t, d_{\text{embed}}) \tag{4.6}$$

where $\tau(t)$ is a d_{embed}-dimensional vector used to encode the timestep information. The function follows the sinusoidal positional encoding approach:

$$\gamma(t) = \left[\cos\left(\frac{t}{M^{\frac{2i}{d_{\text{embed}}}}} \right), \sin\left(\frac{t}{M^{\frac{2i}{d_{\text{embed}}}}} \right) \right]_{i=0}^{d_{\text{embed}}/2} \tag{4.7}$$

where:

- M is a scaling factor (default $10,000$) that controls the frequency range.
- i is the embedding index.

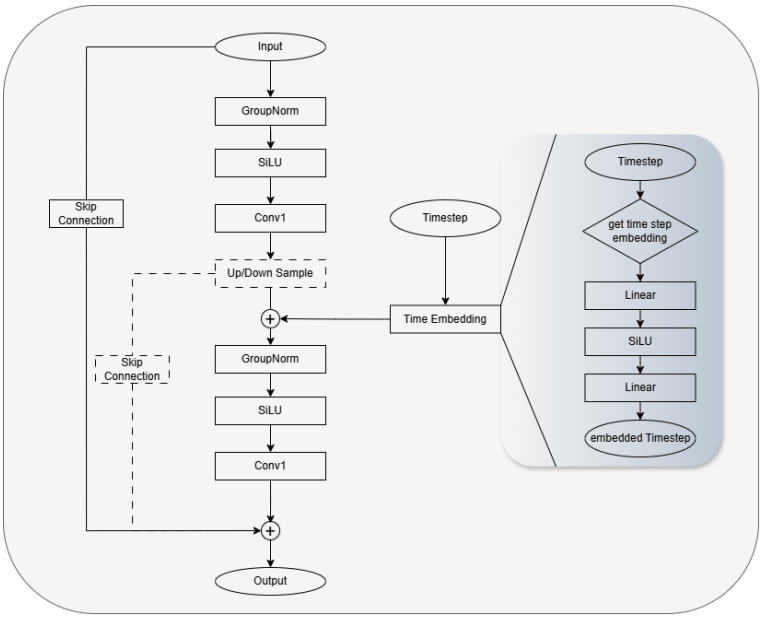


Figure 4.7 – This figure shows how to inject the embedded timestep into the encoder. The dashed lines indicate layers where attention is applied in the final encoder block and the first decoder block.

After obtaining the sinusoidal timestep embedding $\tau(t)$, I further transform it using a multi-layer perceptron (MLP) to project it into a higher-dimensional space. This transformation ensures that the timestep information is effectively integrated into the model.

The embedding dimension is defined as:

$$d_{\text{embed}} = 4 \times d_{\text{channels}} \tag{4.8}$$

where d_{channels} corresponds to the number of feature channels in the first layer of the model. The timestep embedding undergoes the following transformations:

$$\tau'(t) = \text{SiLU}(\text{Linear}(\tau(t), d_{\text{embed}})) \tag{4.9}$$

$$\tau''(t) = \text{Linear}(\tau'(t), d_{\text{embed}}) \tag{4.10}$$

where:

- $\text{Linear}(x, d)$ represents a fully connected layer projecting x to dimension d.
- $\text{SiLU}(x)$ is the Sigmoid-Weighted Linear Unit (SiLU) activation function, which helps to capture complex non-linear dependencies in the timestep embeddings.

The encoded timestep $\tau''(t)$ is injected into each **ResNet block** in the encoder. The way to inject $\tau''(t)$ to encoder is shown in Figure 4.7

4.3 Experiment Setup

4.3.1 Dataset

In this study, we use the data from Ultra-low Dose PET Imaging Challenge[1]. 387 patient data were acquired from the Siemens Biograph Vision Quadra scanner. Images are reconstructed with OSEM to be $440 \times 440 \times 644$ voxels at a voxel spacing of $1.65mm^3$.For Full-TrSUN, we allocate 75 patients for testing and 30 patients for validation. For the further time-controlled TCCA-Net, we allocate 50 patients for testing and 50 patients for validation. All acquired data was in 'list mode', which can be used to reconfigure the acquisition duration to simulate various low-dose scenarios. Each low-dose PET was characterized by a DRF derived from reconstructed counts over a reduced acquisition time frame centered around the midpoint of the original

acquisition duration. The generation of low-dose images employed DRFs of 2, 4, 10, 20, 50, and 100, in conjunction with a corresponding full-dose image, to span a comprehensive range of dose levels. Our research focuses on minimizing radiation exposure by synthesizing high-quality standard-dose images from the lowest dose inputs; hence, we exclusively utilize the DRF of 100, representing a dose reduction of 100 percent, for our training input.

4.3.2 Implementation Details

All experiments were conducted using PyTorch with TensorBoard for visual analytics, using i7-5930K CPU and 24 GB NVIDIA RTX A3090 GPU. The learning rate was set to 0.0002 and batch size was set to 1. We trained our method for more than 300,000 iterations.

4.3.3 Experimental Settings

The study focused on the ultra-low-dose synthesis and therefore only used DRF100 as the input. For testing, to demonstrate the enhanced generalizability of our model, we will evaluate it on PET image data across various dose levels. After comparing different model parameters empirically (in Section 4.4), we found that a larger dimensionality (*dim*) of the Transformer block leads to improved results. Here, we set the Full-TrSUN dimension to 16, matching that of TCCA-Net to ensure a fair comparison. The architecture of the Transformer stage was set to $(1, 2, 2, 2, 1)$, and the number of attention heads was set to $(1, 2, 4, 8, 16)$. The number of channels for the Resnet encoder was set to $(16, 32, 64, 128)$. Additionally, for the hidden states encoder and the UNet skip-connection for the decoding processes, the $feature\_size$ was set to 16, and the *spatial_dims* parameter was set to 3.

For Full-TrSUN and AE-ResUNet, we use the L1 loss to minimize the pixel-wise difference between generated and target images:

$$\mathcal{L}_{L1} = \frac{1}{N} \sum |X_{\text{pred}} - X_{\text{target}}| \tag{4.11}$$

In the time-controlled strategy, two losses are applied:

- SPET Image Loss (L1 Loss): Ensures accurate reconstruction of the high-quality SPET image:

$$\mathcal{L}_{SPET} = \frac{1}{N} \sum |X_{\text{SPET}}^{\text{pred}} - X_{\text{SPET}}^{\text{target}}| \tag{4.12}$$

- Noisy PET Image Loss (MSE Loss): Measures the discrepancy between the generated noisy PET image X_t and the LPET image:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum (X_t - X_{\text{LPET}})^2 \tag{4.13}$$

The total loss function is:

$$\mathcal{L} = \mathcal{L}_{SPET} + \mathcal{L}_{\text{MSE}} \tag{4.14}$$

This formulation ensures both accurate SPET synthesize and robust noise modeling.

We evaluate the effectiveness of the synthesized results using the commonly used evaluation metrics, including PSNR, SSIM, NRMSE and MSE.

4.3.4 Comparison Methods

The Full-TrSUN and time-controlled TCCA-Net model was evaluated against several state-of-the-art synthesis methods. For comparison, we included the traditional 3D-UNet [11], four GAN-based methods specifically designed for PET synthesis—cGAN [79], Cycle-GAN [95, 96], AR-GAN [46], and SS-AEGAN [85]—as well as two transformer-based methods, Restormer [90] and SPACH-Transformer [35]. To ensure a fair comparison, all models were implemented using PyTorch. Due to GPU memory limita-

tions, the dimensional settings for Restormer and SPACH-Transformer were adjusted to 6 and 10, respectively.

4.4 Results and Discussion

4.4.1 Full-TrSUN Results on DRF100

Comparison with the State-of-the-Art (SOTA) models

The results of the proposed Full-TrSUN model and various comparative methods are summarized in Table 4.1. The visualisation results in Figure 4.8 further demonstrate that the Full-TrSUN model performs the best among the comparison methods.

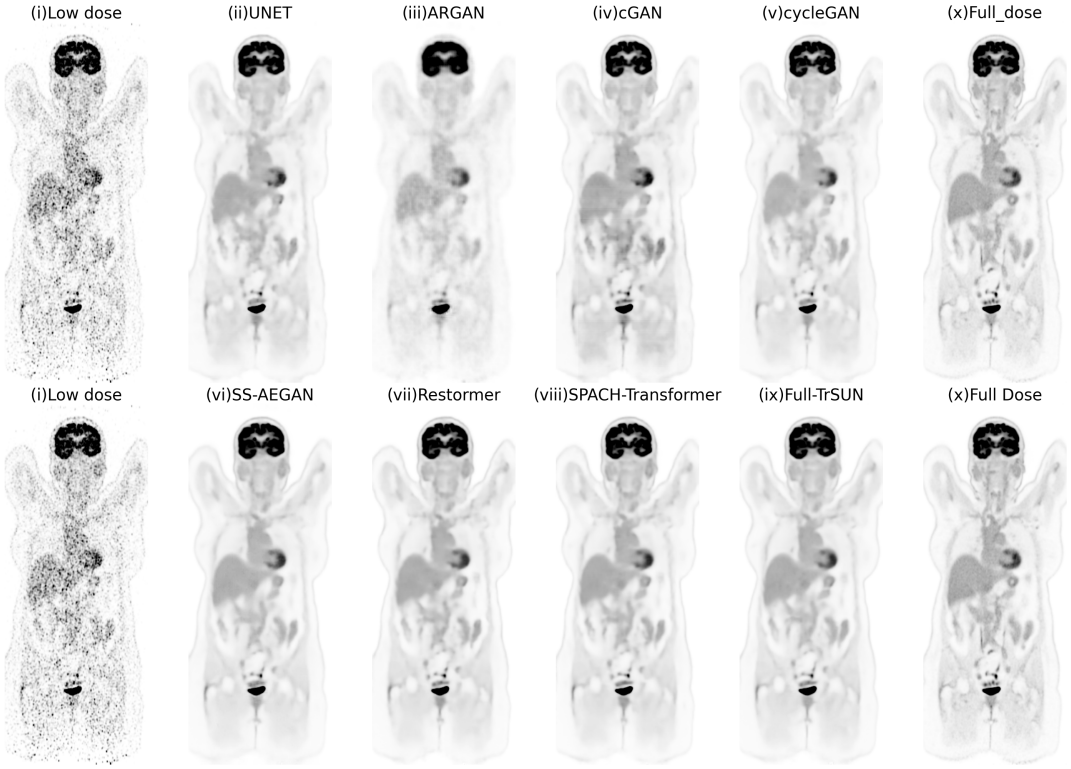


Figure 4.8 – Visualization of different models.

The boundaries in Full-TrSUN are clearer then the other methods. The Full-TrSUN

method consistently surpassed other methods across all evaluation metrics, suggesting that it is the most effective technique for synthesizing SPET images from LPET images. The Full-TrSUN model achieved highest PSNR of 52.7788 dB, indicating that the synthesized images exhibit higher fidelity and closer resemblance to the ground truth compared to those produced by comparison models. Additionally, with an SSIM of 0.9927, the Full-TrSUN model resulted in superior structural similarity to the original SPET images, reflecting its capability to accurately preserve image textures and structures. Furthermore, the model recorded the lowest MSE of 0.0296, suggesting minimal errors in image reconstruction and high precision in replicating the original images. The lowest NRMSE of 0.2733% by Full-TrSUN model further underscores its effectiveness in minimizing the discrepancy between the synthesized and original images.

Table 4.1 – Comparison with the State of the Art models for Full-TrSUN. In this table, bold indicates the best result, and underline indicates the second-best result.

	PSNR(dB)(↑)	SSIM(↑)	MSE(↓)	NRMSE(%)(↓)	Memory(GB)
raw	42.6927	0.9141	0.3235	0.8992	-
unet	49.8706	0.9887	0.0665	0.3770	4.0
argan	47.4489	0.9782	0.1046	0.4809	5.1
cgan	49.1875	0.9876	0.0636	0.4265	4.3
cyclegan	50.6001	0.9907	0.0471	0.3509	10.5
SS-AEGAN	52.0535	0.9918	0.0340	0.2943	4.5
Restormer	52.0575	0.9874	0.0346	0.2958	19.7
SPACH-Transformer	<u>52.1662</u>	<u>0.9857</u>	<u>0.0339</u>	<u>0.2926</u>	20.2
Full Tr-SUN	**52.7788**	**0.9927**	**0.0296**	**0.2733**	21.7

Comparison of model parameters

This section further evaluated the impact of varying the *number of blocks* and the parameter *dim* on the model's performance.

Table 4.2 presents the results for different block configurations with *dim* set to 12. The configuration $[1, 2, 2, 2, 1]$ demonstrated the best performance, achieving a PSNR of 51.9578 dB, an SSIM of 0.9915, an NRMSE of 0.2980%, and an MSE of 0.0381. These

Table 4.2 – Comparison of parameter num_block with ten epochs training using dim=12. In this table, bold indicates the best result, and underline indicates the second-best result.

num_blocks	PSNR(dB)(↑)	SSIM(↑)	NRMSE(%)(↓)	MSE(↓)
[4, 6, 6, 6, 4]	51.6793	0.9916	0.3058	0.0489
[2, 3, 3, 3, 2]	51.8052	0.9909	0.3024	0.0423
[2, 2, 2, 2, 2]	51.6094	0.9910	0.3082	0.0437
[1, 2, 2, 2, 1]	**51.9578**	**0.9915**	**0.2980**	**0.0381**
[1, 1, 1, 1, 1]	<u>51.8097</u>	<u>0.9908</u>	<u>0.3027</u>	<u>0.0424</u>

Table 4.3 – Comparison of parameter dim with ten epochs training using num_block = (1,2,2,2,1). In this table, bold indicates the best result, and underline indicates the second-best result.

dim	PSNR(dB)(↑)	SSIM(↑)	NRMSE(%)(↓)	MSE(↓)
6	51.2461	0.9882	0.3191	0.0520
12	<u>51.9578</u>	**0.9915**	<u>0.2980</u>	<u>0.0381</u>
24	**52.0575**	<u>0.9874</u>	**0.2958**	**0.0346**

metrics suggest that this configuration effectively balances computational efficiency and model performance.

The effect of varying *dim* values was assessed, starting with the optimal block configuration $[1, 2, 2, 2, 1]$. Table 4.3 shows that under limited GPU memory constraints, a *dim* value of 24 yielded the best results, achieving a PSNR of 52.0575 dB, an SSIM of 0.9874, an NRMSE of 0.2958%, and an MSE of 0.0346. This larger dim value enhances the model's capacity to capture more complex spatial relationships, thus improving image quality and reconstruction accuracy. In the context of channel-wise transformer methods, the parameter ***dim*** is particularly critical. Under the same GPU constraints, this Full-TrSUN model can utilize a larger ***dim*** value. Using a CNN encoder-decoder structure to replace the transformer decoder process can help save memory, enabling the use of better parameter settings to achieve higher results. This approach leverages the spatial relationship capabilities of CNNs, enhancing computational efficiency and performance.

4.4.2 Performance Comparison Across DRFs

To evaluate the effectiveness of Full-TrSUN and time-controlled TCCA-Net, comparison of the performance across different DRFs are condcuted using the evaluation metrics of PSNR, SSIM, MSE, and NRMSE. The results show in Table 4.4, where different models perform optimally at different noise levels, highlight the importance of adaptive reconstruction strategies. The visualization of example patient is shown in Figure 4.9 and Figure 4.10. These two figures provide visual comparisons of PET image synthesis across different dose levels using various state-of-the-art (SOTA) methods, including Full-TrSUN and Time-Controlled TCCA-Net. Figure 4.9 presents the visualization of an example patient synthesized by different models at DRF100, DRF50, and DRF20. Figure 4.10 presents the visualization of an example patient synthesized by different models at DRF10, DRF4, and DRF2.

At the lower dose level DRF100 and DRF50, Full-TrSUN achieved the best performance. The model preserved structural details more effectively than the comparison methods, resulting in the highest PSNR and SSIM while maintaining the lowest MSE and NRMSE. It primarily due to its hierarchical transformer architecture, which effectively extracts detailed multi-scale features directly from full-resolution images. This design allows Full-TrSUN to better capture subtle textural nuances and spatial details, which are particularly critical at very low-dose levels characterized by higher noise and fewer distinguishable features. In contrast, the time-controlled method primarily focuses on noise-level modeling, offering adaptability across various dose conditions but potentially sacrificing fine-grained feature representation at extremely low doses.

As the dose level increases, resulting in clearer images (e.g., DRF20 and DRF10), the time-controlled TCCA-Net method outperformed Full-TrSUN. This expected improvement is due to the TCCA-Net's explicit modeling of noise variations through its time-controlled mechanism, which provides continuous, adaptive guidance during image synthesis. Additionally, the attention-enhanced ResUNet architecture integrates spatial and channel-wise attention mechanisms, effectively highlighting relevant im-

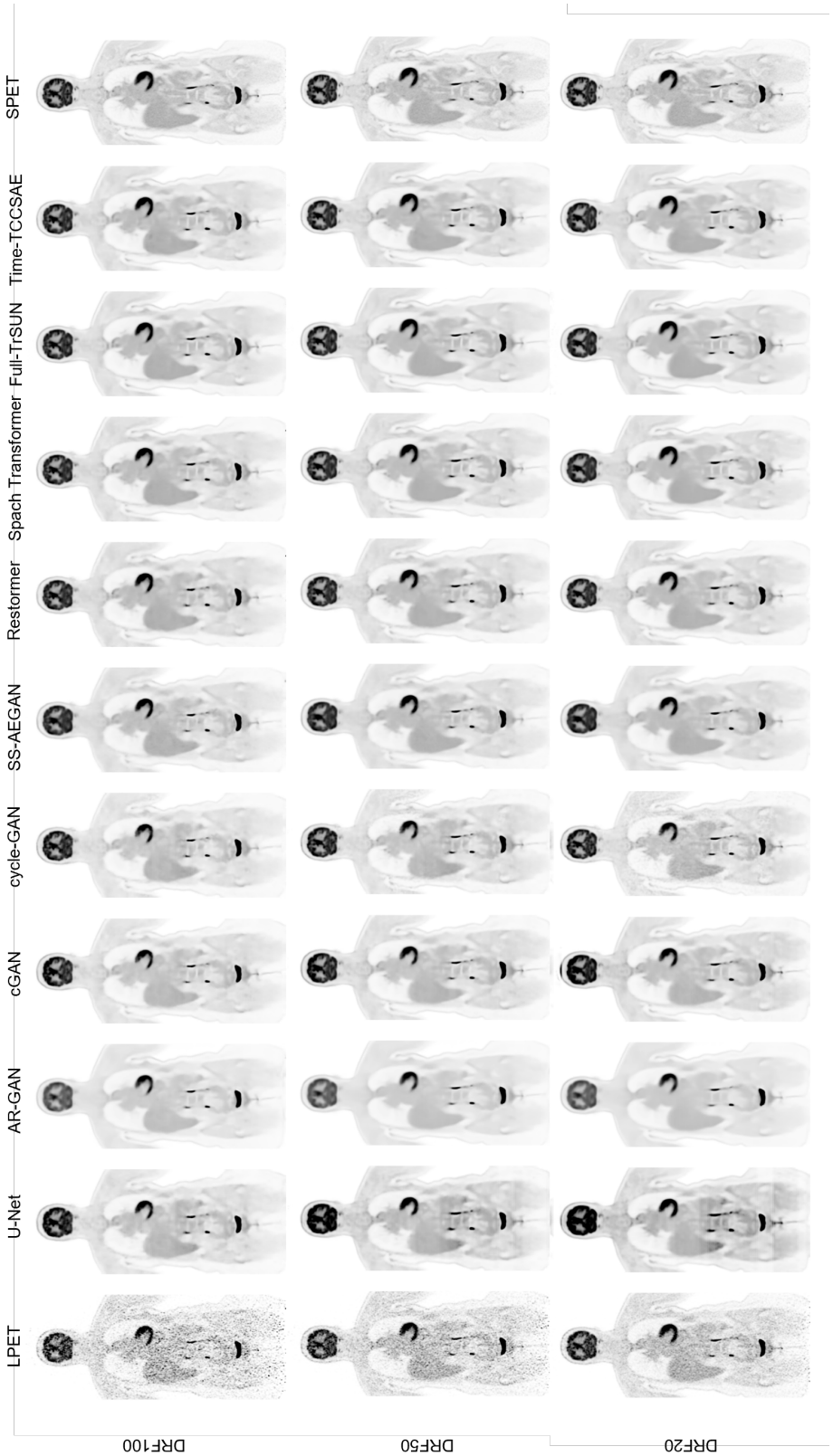


Figure 4.9 – Comparison of PET image reconstruction across different dose levels.

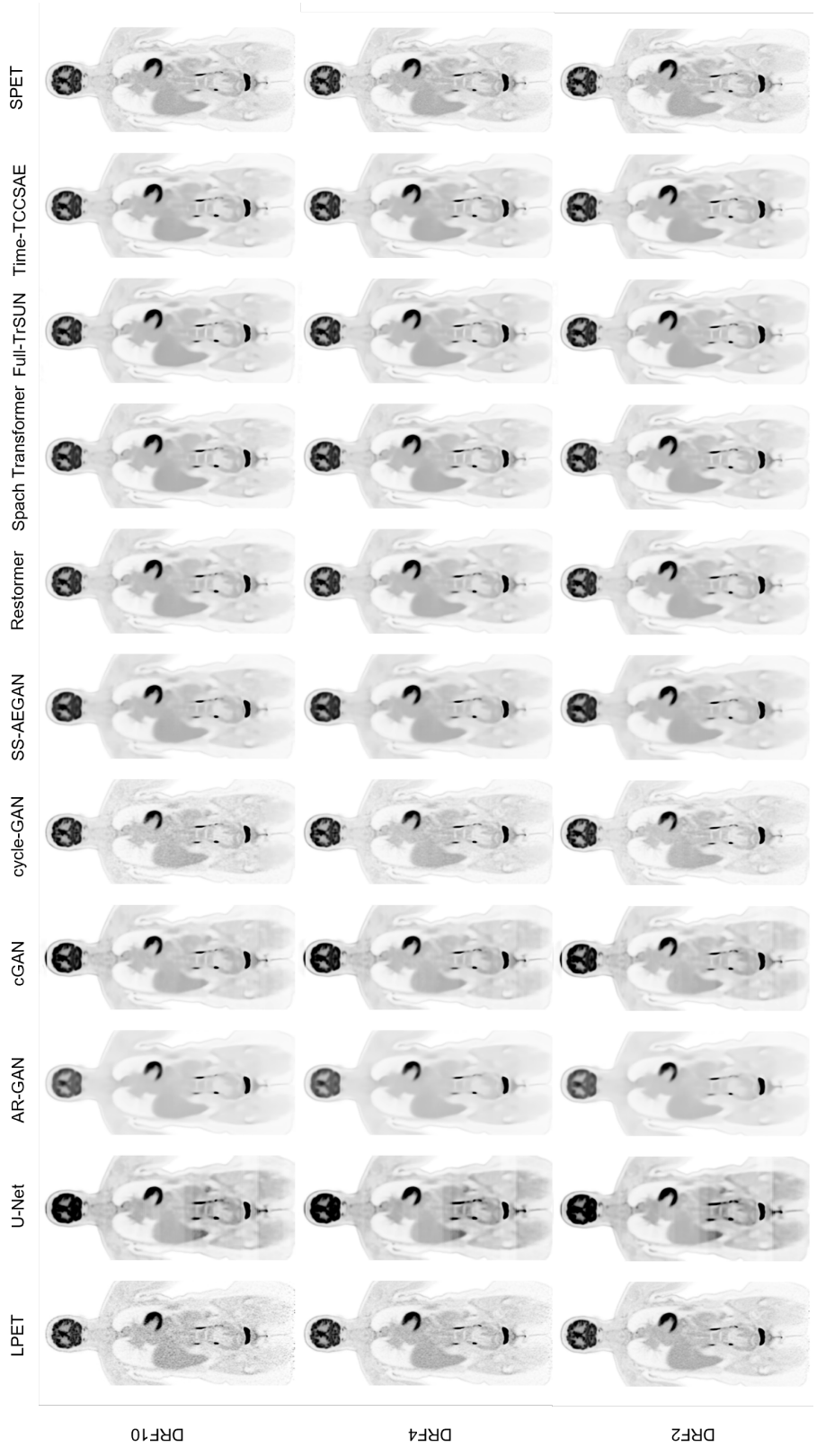


Figure 4.10 – Comparison of PET image reconstruction across different dose levels.

Table 4.4 – Comparison with the State of the Art models for TCCA-Net w/o time. In this table, bold indicates the best result, and underline indicates the second-best result.

		RAW	UNet	AR-GAN	cGAN	cycleGAN	SS-AEGAN	Restormer	SPACH Transformer	TRSUN	TCCA-Net
DRF100	PSNR	42.6002	49.8146	47.1175	50.8217	50.3326	51.3726	52.0153	52.3746	**52.7905**	52.5963
	SSIM	0.9092	0.9881	0.9901	0.9899	0.9894	0.9903	0.9830	0.9887	**0.9931**	0.9910
	MSE	0.3245	0.0532	0.0987	0.0425	0.0484	0.0384	0.0336	0.0310	**0.0284**	0.0291
	NRMSE	0.9024%	0.3875%	0.5048%	0.3350%	0.3550%	0.3125%	0.2930%	0.2821%	**0.2683%**	0.2755%
DRF50	PSNR	46.3044	48.9994	47.6826	51.5279	51.3676	52.5508	53.6811	54.0136	**54.1589**	54.1354
	SSIM	0.9396	0.9905	0.9891	0.9930	0.9918	0.9911	0.9925	0.9945	**0.9947**	0.9934
	MSE	0.1240	0.0630	0.0866	0.0346	0.0372	0.0285	0.0220	0.0204	**0.0203**	0.0199
	NRMSE	0.5788%	0.4458%	0.4716%	0.3110%	0.3178%	0.2703%	0.2403%	0.2322%	**0.2275%**	0.2290%
DRF20	PSNR	50.5299	46.6952	47.9060	50.6380	51.7825	53.2936	55.1423	55.3226	55.0245	**55.5471**
	SSIM	0.9783	0.9891	0.9888	0.9915	0.9923	0.9916	0.9948	**0.9958**	0.9952	0.9951
	MSE	0.0450	0.1067	0.0825	0.0430	0.0333	0.0240	0.0157	0.0149	0.0165	**0.0144**
	NRMSE	0.3523%	0.5876%	0.4589%	0.3587%	0.3067%	0.2470%	0.2013%	0.1987%	0.2054%	**0.1926%**
DRF10	PSNR	53.4283	45.6202	47.9678	49.3415	52.0056	53.5724	55.9259	56.0875	55.5303	**56.2705**
	SSIM	0.9925	0.9873	0.9887	0.9876	0.9927	0.9917	0.9953	**0.9962**	0.9954	0.9956
	MSE	0.0231	0.1353	0.0813	0.0599	0.0317	0.0224	0.0131	0.0124	0.0146	**0.0121**
	NRMSE	0.2504%	0.6635%	0.4553%	0.4300%	0.2997%	0.2384%	0.1825%	0.1809%	0.1931%	**0.1755%**
DRF4	PSNR	57.4371	44.9269	48.0193	48.1457	52.4818	53.8472	56.8054	**57.1256**	56.1468	57.0682
	SSIM	0.9978	0.9870	0.9884	0.9843	0.9934	0.9916	0.9957	**0.9967**	0.9955	0.9959
	MSE	0.0092	0.1576	0.0802	0.0840	0.0296	0.0208	0.0106	**0.0097**	0.0126	0.0099
	NRMSE	0.1741%	0.7158%	0.4521%	0.5067%	0.2860%	0.2304%	0.1634%	0.1594%	0.1798%	**0.1586%**
DRF2	PSNR	61.7018	44.4495	48.0352	47.4198	52.9215	54.0389	57.4961	**58.0311**	56.6748	57.6942
	SSIM	0.9992	0.9861	0.9882	0.9850	0.9945	0.9914	0.9958	**0.9970**	0.9954	0.9961
	MSE	0.0035	0.1824	0.0798	0.1111	0.0306	0.0199	0.0090	**0.0078**	0.0111	0.0085
	NRMSE	0.0947%	0.7632%	0.4514%	0.5513%	0.2780%	0.2252%	0.1505%	**0.1434%**	0.1695%	0.1469%

age features and suppressing redundant information. Together, these components significantly enhance feature clarity and detail at higher dose levels, where precise feature representation becomes more critical than noise modeling alone.

At DRF4 and DRF2, where input images have minimal noise, the SPACH-Transformer achieved superior overall performance. This is likely attributed to its pure transformer-based architecture, integrating both spatial and channel-wise attention, and thus excelling at modeling long-range dependencies and capturing global contextual features. In scenarios with minimal noise interference, the primary challenge shifts toward effectively capturing subtle global context and intricate details rather than noise reduction. Under these conditions, the SPACH-Transformer's ability to leverage comprehensive spatial-channel attention without convolutional constraints becomes particularly advantageous, enabling it to deliver the best synthesis quality. At these dose levels (DRF4 and DRF2), although TCCA-Net did not achieve the highest performance, it still achieved second-best results. This indicates that the TCCA-Net, leveraging its time-controlled mechanism and attention-enhanced ResUNet structure, maintained robust and consistent performance across varying dose conditions. Its adaptability and continuous noise-level modeling ensured it remained competitive even when the primary focus shifts from noise reduction to fine-grained global feature extraction.

Overall, the presented results suggest that Full-TrSUN is optimal for highly noisy PET images (DRF100 and DRF50), as its hierarchical transformer was able to effectively enhance the synthesizes accuracy under challenging conditions. As noise decreases (DRF20 and DRF10), TCCA-Net surpasses Full-TrSUN due to its adaptive encoding of noise variations, improving performance across varying dose levels. At the lowest noise scenarios (DRF4 and DRF2), SPACH-Transformer achieves the best overall results; however, TCCA-Net remains a strong second-best, with only marginal differences in performance (approximately), highlighting TCCA-Net's consistent intensity preservation and robust PET image synthesis capabilities. At DRF4, SPACH slightly outperforms TCCA-Net, achieving approximately 0.06 higher PSNR, 0.0008 higher SSIM, and 0.0002 lower MSE. However, TCCA-Net demonstrates a marginally better NRMSE, lower by 0.0008%, indicating enhanced intensity consistency. Similarly,

at DRF2, SPACH maintains a slight advantage over TCCA-Net, with PSNR higher by about 0.04, SSIM improved by 0.0009, and MSE reduced by 0.0007. In terms of NRMSE, SPACH remains marginally better, lower by 0.0035%.

4.4.3 Ablation Study

To evaluate the influence of key model components, an ablation study is conducted across four configurations: (1) Full-TrSUN (channel-wise attention only); (2) Full-TrSUN combined with ResUNet, featuring two encoders—one with channel-wise attention and one standard ResUNet encoder without attention; (3) TCCA-Net without time-controlled encoding, comprising two encoders—one with channel-wise attention and another ResUNet encoder integrating spatial-wise attention at the final encoder block and the first decoder layer; and (4) the complete time-controlled TCCA-Net model, which includes the dual-encoder structure (channel-wise and ResUNet with spatial-wise attention) and incorporates the time-controlled mechanism. The study specifically investigates how the addition of attention mechanisms and time-controlled encoding impacts reconstruction quality across various dose reduction factors (DRFs).

At the highest noise level, DRF100, Full-TrSUN achieves the best performance, demonstrating its ability to reconstruct PET images under extreme noise conditions. The transformer-based attention mechanisms in Full-TrSUN effectively enhance feature selection and structural preservation, leading to the highest PSNR and SSIM while minimizing MSE and NRMSE. Full-TrSUN also performs best at DRF50, maintaining its advantage when the input images still contain significant noise.

As the noise level decreases (DRF20 and lower), TCCA-Net surpasses Full-TrSUN, indicating that its architecture is more effective in handling moderately noisy PET images. While Full-TrSUN benefits from attention mechanisms at high noise levels, its performance declines as the input quality improves. In contrast, TCCA-Net provides better reconstruction results, likely due to its enhanced feature representation tailored for moderate-dose conditions.

Across all DRFs, the addition of time-controlled encoding consistently improves per-

Table 4.5 – Albation study of TCCA-Net. In this table, bold indicates the best result, and underline indicates the second-best result.

		RAW	Full-TrSUN	Full-TrSUN +ResUNet	AE-ResUNet	TCCA-Net
DRF100	PSNR	42.6002	**52.7905**	52.2557	52.5043	<u>52.5963</u>
	SSIM	0.9092	**0.9931**	0.9904	<u>0.9917</u>	0.9910
	MSE	0.3245	**0.0284**	0.0316	0.0297	<u>0.0291</u>
	NRMSE	0.9024%	**0.2683%**	0.2845%	0.2782%	<u>0.2755%</u>
DRF50	PSNR	46.3044	**54.1589**	53.7488	53.9965	<u>54.1354</u>
	SSIM	0.9396	**0.9947**	0.9933	0.9939	<u>0.9934</u>
	MSE	0.1240	<u>0.0203</u>	0.0220	0.0205	**0.0199**
	NRMSE	0.5788%	**0.2275%**	0.2376%	0.2324%	<u>0.2290%</u>
DRF20	PSNR	50.5299	55.0245	55.0240	<u>55.2670</u>	**55.5471**
	SSIM	0.9783	0.9952	**0.9954**	**0.9954**	0.9951
	MSE	0.0450	0.0165	<u>0.0152</u>	<u>0.0152</u>	**0.0144**
	NRMSE	0.3523%	0.2054%	<u>0.1987%</u>	<u>0.1987%</u>	**0.1926%**
DRF10	PSNR	53.4283	55.5303	55.6583	<u>55.9074</u>	**56.2705**
	SSIM	0.9925	0.9954	0.9954	**0.9958**	<u>0.9956</u>
	MSE	0.0231	0.0146	0.0142	<u>0.0131</u>	**0.0121**
	NRMSE	0.2504%	0.1931%	0.1874%	<u>0.1831%</u>	**0.1755%**
DRF4	PSNR	57.4371	56.1468	56.3297	<u>56.6192</u>	**57.0682**
	SSIM	0.9978	0.9955	0.9957	**0.9960**	0.9959
	MSE	0.0092	0.0126	0.0121	<u>0.0109</u>	**0.0099**
	NRMSE	0.1741%	0.1798%	0.1719%	<u>0.1673%</u>	**0.1586%**
DRF2	PSNR	61.7018	56.6748	56.8440	<u>57.1700</u>	**57.6942**
	SSIM	0.9992	0.9954	0.9958	**0.9962**	<u>0.9961</u>
	MSE	0.0035	0.0111	0.0108	<u>0.0096</u>	**0.0085**
	NRMSE	0.0947%	0.1695%	<u>0.1565%</u>	<u>0.1565%</u>	**0.1469%**

formance over the standard TCCA-Net model. Time-controlled TCCA-Net achieves superior PSNR and SSIM while reducing MSE and NRMSE at every dose level. The improvement is most significant at DRF20 and DRF10, where time-controlled encoding enhances generalization and allows the model to adapt to different noise conditions effectively.

These findings confirm that Full-TrSUN is optimal for extremely noisy PET images (DRF100 and DRF50), where attention mechanisms improve reconstruction quality. However, as noise levels decrease, TCCA-Net becomes more effective than Full-TrSUN, and with time-controlled encoding, time-controlled TCCA-Net consistently achieves the best results across all DRFs. The ability of time-controlled TCCA-Net to maintain lower NRMSE at DRF4 highlights its robustness in preserving structural details and intensity consistency, making it particularly effective for low-dose PET image synthesis across varying noise conditions.

4.5 Summary

This chapter introduced Time-Controlled TCCA-Net, an improved model based on Full-TrSUN, designed for PET image synthesize across varying dose levels. The time-controlled mechanism enhances noise adaptation, enabling dynamic adjustment to different input conditions.

The performance comparison highlights that model selection should be tailored to the noise level of PET images for optimal results. Full-TrSUN performs best under high-noise conditions (e.g., DRF100 and DRF50), where its transformer-based attention mechanisms enhance feature selection and structural preservation, resulting in superior PSNR and SSIM and reduced MSE and NRMSE. As noise levels decrease (e.g., DRF20 and below), Full-TrSUN's performance declines, while TCCA-Net with time-controlled encoding begins to outperform it, demonstrating strong generalization and achieving the lowest NRMSE at DRF4.

At extremely low-noise levels such as DRF4 and DRF2, SPACH slightly outperforms

TCCA-Net, suggesting its architecture is better suited for near-clean inputs where fine detail preservation is crucial. However, SPACH performs poorly under high-noise conditions, showing a significant drop in synthesize quality. In contrast, models like TrSUN exhibit inconsistent performance across noise levels. These results underscore the importance of noise-adaptive model design and dose-aware training strategies to ensure robust synthesis across varying PET imaging conditions.

For future work, a more generalizable model will be developed that can adapt to all the dose levels without requiring separate training. Enhancing time-controlled encoding, incorporating adaptive learning strategies, and combining Transformer-based global attention with CNN-based local feature extraction are potential directions. Optimizing computational efficiency will further enable real-world deployment.

Chapter 5

Conclusions and Future Work

5.1 Conclusion

The work in this thesis presents a data augmentation strategy and a time-controlled framework for high-quality PET image synthesis, addressing the challenge of generating accurate reconstructions from LPET images. A reverse method for data augmentation was proposed, leveraging routine SPET images to generate synthetic LPET counterparts, increasing training data diversity and improving PET image synthesis performance. Additionally, a novel deep learning architecture was introduced, integrating channel-wise attention, ResNet blocks, and spatial attention, along with a time-controlled mechanism to improve adaptability across different noise conditions.

The reverse data augmentation strategy played a crucial role in mitigating the challenge of limited training data for PET synthesis models. By generating synthetic LPET images from SPET images, this method expanded the training dataset, improving model robustness and generalization across varying dose levels. The reverse method was particularly effective in enhancing the model's ability to synthesize high-quality PET images with improved PSNR, SSIM, and lower MSE. These results confirm that integrating time-controlled encoding and reverse augmentation techniques significantly improves PET image reconstruction, making deep learning models more adaptable to real-world clinical scenarios.

Furthermore, the experimental results demonstrated that Full-TrSUN achieves the best performance at high noise levels (DRF100, DRF50), while TCCA-Net outperforms Full-TrSUN at lower noise levels (DRF20 and below), demonstrating its superior adaptability. At DRF4, TCCA-Net achieves the lowest NRMSE, confirming its effectiveness in maintaining intensity consistency. The ablation study validated the impact of attention mechanisms in Full-TrSUN for feature selection in noisy conditions and the time-controlled encoding in TCCA-Net, which consistently improved reconstruction quality across all dose levels.

5.2 Future Work

While the proposed methods achieved state-of-the-art performance, several areas for improvement remain. A key direction for future research is to explore the integration of our data augmentation strategy with the time-controlled framework, aiming to further enhance its adaptability across varying dose levels. While the time-controlled model already demonstrates strong generalization, combining it with tailored augmentation techniques could further improve its robustness, potentially enabling a single model to effectively handle all noise conditions without requiring separate training for each dose level. This could be achieved by enhancing the time-controlled encoding mechanism, allowing the model to better capture noise distribution patterns. Additionally, self-supervised learning techniques could be integrated to reduce dependence on annotated datasets, improving scalability for clinical applications.

Further refinement of the reverse data augmentation method can be explored to improve the augmented image quality. Currently, LPET images generated from SPET images provide effective data augmentation, but improving the accuracy of these synthetic LPET images is essential. Future work could explore advanced generative models such as diffusion models or domain adaptation techniques to enhance the realism of synthetic LPET images. Moreover, uncertainty estimation techniques could be incorporated to assess the confidence of generated LPET images, ensuring more reliable training data.

Beyond architectural improvements, computational efficiency remains a crucial factor for real-world deployment. Optimizing inference speed is essential for real-time or near real-time PET image synthesis, ensuring seamless integration into clinical workflows. Furthermore, large-scale clinical validation using diverse PET datasets is necessary to evaluate the model's reliability and effectiveness. Collaborations with medical professionals can help define clinically relevant evaluation metrics, ensuring that improvements in image quality align with diagnostic accuracy and clinical usability.

Furthermore, aligning with recent advancements in medical AI, future research could explore multimodal learning frameworks that integrate PET with complementary modalities such as CT or MRI. Such multimodal enhancement has the potential to leverage structural information from CT/MRI to further improve the quality and clinical value of synthesized PET images. The use of large-scale pretrained models or foundation models could also be investigated to improve transferability and representation learning, especially in low-data scenarios. These models may offer better generalization across imaging domains and reduce the need for modality-specific training from scratch.

In summary, future research should focus on developing a fully dose-adaptive PET reconstruction framework that generalizes across all dose levels, enhances computational efficiency, and undergoes clinical validation for real-world applicability. By addressing these challenges, deep learning-based PET imaging can be further advanced, improving the accuracy and accessibility of medical imaging for broader clinical use.

References

[1] (2023). Ultra-low dose pet imaging challenge.

[2] Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al. (2021). Xcit: Cross-covariance image transformers. *Advances in neural information processing systems*, 34:20014–20027.

[3] Antoniou, A., Storkey, A., and Edwards, H. (2017). Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*.

[4] Bailey, D. L., Maisey, M. N., Townsend, D. W., and Valk, P. E. (2005). *Positron emission tomography*, volume 2. Springer.

[5] Bankman, I. (2008). *Handbook of medical image processing and analysis*. Elsevier.

[6] Ben-Cohen, A., Klang, E., Raskin, S. P., Soffer, S., Ben-Haim, S., Konen, E., Amitai, M. M., and Greenspan, H. (2019). Cross-modality synthesis from ct to pet using fcn and gan networks for improved automated lesion detection. *Engineering Applications of Artificial Intelligence*, 78:186–194.

[7] Bi, L., Kim, J., Kumar, A., Feng, D., and Fulham, M. (2017). Synthesis of positron emission tomography (pet) images via multi-channel generative adversarial networks (gans). In *Molecular Imaging, Reconstruction and Analysis of Moving Body Organs, and Stroke Imaging and Treatment: Fifth International Workshop, CMMI 2017, Second International Workshop, RAMBO 2017, and First International Workshop, SWITCH 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, 2017, Proceedings 5*, pages 43–51. Springer.

[8] Chaitanya, K., Karani, N., Baumgartner, C. F., Erdil, E., Becker, A., Donati, O., and Konukoglu, E. (2021). Semi-supervised task-driven data augmentation for medical image segmentation. *Medical Image Analysis*, 68:101934.

[9] Cherry, S. R., Sorenson, J. A., and Phelps, M. E. (2013). *Physics in nuclear medicine*. Saunders.

[10] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563.

[11] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*, pages 424–432. Springer.

[12] Cui, J., Gong, K., Guo, N., Wu, C., Meng, X., Kim, K., Zheng, K., Wu, Z., Fu, L., Xu, B., et al. (2019). Pet image denoising using unsupervised deep learning. *European journal of nuclear medicine and molecular imaging*, 46:2780–2789.

[13] Cui, J., Wang, Y., Zhou, L., Fei, Y., Zhou, J., and Shen, D. (2024a). 3d point-based multi-modal context clusters gan for low-dose pet image denoising. *IEEE Transactions on Circuits and Systems for Video Technology*.

[14] Cui, J., Zeng, P., Xu, Y., Wu, X., Zhou, J., and Wang, Y. (2024b). S3pet: Semi-supervised standard-dose pet image reconstruction via dose-aware token swap. In *2024 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE International Conference on Robotics, Automation and Mechatronics (RAM)*, pages 20–25. IEEE.

[15] Cui, J., Zeng, P., Zeng, X., Wang, P., Wu, X., Zhou, J., Wang, Y., and Shen, D. (2023). Trido-former: A triple-domain transformer for direct pet reconstruction from low-dose sinograms. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 184–194. Springer.

[16] Cui, J., Zeng, P., Zeng, X., Xu, Y., Wang, P., Zhou, J., Wang, Y., and Shen, D. (2024c). Prior knowledge-guided triple-domain transformer-gan for direct pet reconstruction from low-count sinograms. *IEEE Transactions on Medical Imaging*.

[17] Cui, J., Zeng, X., Zeng, P., Liu, B., Wu, X., Zhou, J., and Wang, Y. (2024d). Mcad: Multi-modal conditioned adversarial diffusion model for high-quality pet image reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 467–477. Springer.

[18] El Jiani, L., El Filali, S., et al. (2022). Overcome medical image data scarcity by data augmentation techniques: A review. In *2022 International Conference on Microelectronics (ICM)*, pages 21–24. IEEE.

[19] Fei, Y., Zu, C., Jiao, Z., Wu, X., Zhou, J., Shen, D., and Wang, Y. (2022). Classification-aided high-quality pet image synthesis via bidirectional contrastive gan with shared information maximization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 527–537. Springer.

[20] Floyd, C. (1991). An artificial neural network for spect image reconstruction. *IEEE transactions on medical imaging*, 10(3):485–487.

[21] Gong, K., Guan, J., Liu, C.-C., and Qi, J. (2018). Pet image denoising using a deep neural network through fine tuning. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 3(2):153–161.

[22] Gong, K., Johnson, K., El Fakhri, G., Li, Q., and Pan, T. (2024). Pet image denoising based on denoising diffusion probabilistic model. *European Journal of Nuclear Medicine and Molecular Imaging*, 51(2):358–368.

[23] Gong, Y., Shan, H., Teng, Y., Tu, N., Li, M., Liang, G., Wang, G., and Wang, S. (2020). Parameter-transferred wasserstein generative adversarial network (pt-wgan) for low-dose pet image denoising. *IEEE transactions on radiation and plasma medical sciences*, 5(2):213–223.

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

[25] Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

[26] Häggström, I., Schmidtlein, C. R., Campanella, G., and Fuchs, T. J. (2019). Deeppet: A deep encoder–decoder network for directly solving the pet image reconstruction inverse problem. *Medical image analysis*, 54:253–262.

[27] Han, Z., Wang, Y., Zhou, L., Wang, P., Yan, B., Zhou, J., Wang, Y., and Shen, D. (2023). Contrastive diffusion model with auxiliary guidance for coarse-to-fine pet reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 239–249. Springer.

[28] Hashimoto, F., Onishi, Y., Ote, K., Tashima, H., Reader, A. J., and Yamaya, T. (2024). Deep learning-based pet image denoising and reconstruction: a review. *Radiological physics and technology*, 17(1):24–46.

[29] Hashimoto, F. and Ote, K. (2024). Reconu-net: a direct pet image reconstruction using u-net architecture with back projection-induced skip connection. *Physics in Medicine & Biology*, 69(10):105022.

[30] Hayat, M. (2009). *Methods of cancer diagnosis, therapy, and prognosis: liver cancer*, volume 5. Springer Science & Business Media.

[31] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

[32] Huang, B., Law, M. W.-M., and Khong, P.-L. (2009). Whole-body pet/ct scanning: estimation of radiation dose and cancer risk. *Radiology*, 251(1):166–174.

[33] Huang, B., Liu, X., Fang, L., Liu, Q., and Li, B. (2024a). Diffusion transformer model with compact prior for low-dose pet reconstruction. *arXiv preprint arXiv:2407.00944*.

[34] Huang, Z., Li, W., Wu, Y., Yang, L., Dong, Y., Yang, Y., Zheng, H., Liang, D., Wang, M., and Hu, Z. (2024b). Accurate whole-brain image enhancement for low-dose integrated pet/mr imaging through spatial brain transformation. *IEEE Journal of Biomedical and Health Informatics*.

[35] Jang, S.-I., Pan, T., Li, Y., Heidari, P., Chen, J., Li, Q., and Gong, K. (2023). Spach transformer: spatial and channel-wise transformer based on local and global self-attentions for pet image denoising. *IEEE transactions on medical imaging*.

[36] Jiang, C., Pan, Y., Cui, Z., Nie, D., and Shen, D. (2023a). Semi-supervised standard-dose pet image generation via region-adaptive normalization and structural consistency constraint. *IEEE transactions on medical imaging*, 42(10):2974–2987.

[37] Jiang, C., Pan, Y., Liu, M., Ma, L., Zhang, X., Liu, J., Xiong, X., and Shen, D. (2023b). Pet-diffusion: Unsupervised pet enhancement based on the latent diffusion model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–12. Springer.

[38] Jiang, C., Pan, Y., and Shen, D. (2023c). Tridornet: Reconstruction of standard-dose pet from low-dose pet in triple (projection, image, and frequency) domains. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

[39] Jiang, C., Tang, Z., Cui, Z., and Shen, D. (2024). Enhancing pet with image generation techniques: Generating standard-dose pet from low-dose pet. In *Generative Machine Learning Models in Medical Image Computing*, pages 209–229. Springer.

[40] Kapoor, V., McCook, B. M., and Torok, F. S. (2004). An introduction to pet-ct imaging. *Radiographics*, 24(2):523–543.

[41] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

[43] Li, Y. and Li, Y. (2024). Petformer network enables ultra-low-dose total-body pet imaging without structural prior. *Physics in Medicine & Biology*, 69(7):075030.

[44] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88.

[45] Luo, Y., Wang, Y., Zu, C., Zhan, B., Wu, X., Zhou, J., Shen, D., and Zhou, L. (2021). 3d transformer-gan for high-quality pet reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VI 24*, pages 276–285. Springer.

[46] Luo, Y., Zhou, L., Zhan, B., Fei, Y., Zhou, J., Wang, Y., and Shen, D. (2022). Adaptive rectification based adversarial network with spectrum constraint for high-quality pet image synthesis. *Medical Image Analysis*, 77:102335.

[47] Lv, Y. and Xi, C. (2021). Pet image reconstruction with deep progressive learning. *Physics in Medicine & Biology*, 66(10):105016.

[48] Maisey, M. N. (2005). Positron emission tomography in clinical medicine. In *Positron Emission Tomography: Basic Sciences*, pages 1–12. Springer.

[49] Matsubara, K., Ibaraki, M., Nemoto, M., Watabe, H., and Kimura, Y. (2022). A review on ai in pet imaging. *Annals of Nuclear Medicine*, 36(2):133–143.

[50] Mehranian, A., Wollenweber, S. D., Walker, M. D., Bradley, K. M., Fielding, P. A., Su, K.-H., Johnsen, R., Kotasidis, F., Jansen, F. P., and McGowan, D. R. (2022). Image enhancement of whole-body oncology [18f]-fdg pet scans using deep neural networks to reduce noise. *European journal of nuclear medicine and molecular imaging*, 49(2):539–549.

[51] Muehllehner, G. and Karp, J. S. (2006). Positron emission tomography. *Physics in Medicine & Biology*, 51(13):R117.

[52] Nguyen, D. T., Nguyen, T. T., Nguyen, H. T., Nguyen, T. T., Pham, H. H., Nguyen, T. H., Truong, T. N., and Nguyen, P. L. (2024). Ct to pet translation: A large-scale dataset and domain-knowledge-guided diffusion approach. *arXiv preprint arXiv:2410.21932*.

[53] on Radiological Protection, I. C. (1977). *Recommendations of the International Commission on Radiological Protection: Adopted January 17, 1977*, volume 1. International Commission on Radiological Protection.

[54] Ouyang, J., Chen, K. T., Gong, E., Pauly, J., and Zaharchuk, G. (2019). Ultra-low-dose pet reconstruction using generative adversarial network with feature matching and task-specific perceptual loss. *Medical physics*, 46(8):3555–3564.

[55] Pain, C. D., Egan, G. F., and Chen, Z. (2022). Deep learning-based image reconstruction and post-processing methods in positron emission tomography for low-dose imaging and resolution enhancement. *European Journal of Nuclear Medicine and Molecular Imaging*, 49(9):3098–3118.

[56] Pan, S., Abouei, E., Peng, J., Qian, J., Wynne, J. F., Wang, T., Chang, C.-W., Roper, J., Nye, J. A., Mao, H., et al. (2023). Full-dose pet synthesis from low-dose pet using high-efficiency diffusion denoising probabilistic model. *arXiv preprint arXiv:2308.13072*.

[57] Pan, S., Abouei, E., Peng, J., Qian, J., Wynne, J. F., Wang, T., Chang, C.-W., Roper, J., Nye, J. A., Mao, H., et al. (2024). Full-dose whole-body pet synthesis from low-dose pet using high-efficiency denoising diffusion probabilistic model: Pet consistency model. *Medical Physics*.

[58] Phase, B. et al. (2006). Health risks from exposure to low levels of ionizing radiation. *Washington, DC: The British Institute of Radiology*.

[59] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer.

[60] Saade, C., Ammous, A., Abi-Ghanem, A. S., Giesel, F., and Asmar, K. (2019). Body weight-based protocols during whole body fdg pet/ct significantly reduces radiation dose without compromising image quality: findings in a large cohort study. *Academic Radiology*, 26(5):658–663.

[61] Sanaat, A., Najafgholizadeh, A., Mazandarani, H. R., and Zaidi, H. (2022). Low dose brain pet imaging using denoising diffusion probabilistic models. In *2022 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–3. IEEE.

[62] Sanaat, A., Shiri, I., Arabi, H., Mainta, I., Nkoulou, R., and Zaidi, H. (2020). Whole-body pet image synthesis from low-dose images using cycle-consistent generative adversarial networks. In *2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC)*, pages 1–3. IEEE.

[63] Sano, A., Nishio, T., Masuda, T., and Karasawa, K. (2021). Denoising pet images for proton therapy using a residual u-net. *Biomedical Physics & Engineering Express*, 7(2):025014.

[64] Schaefferkoetter, J., Yan, J., Ortega, C., Sertic, A., Lechtman, E., Eshet, Y., Metser, U., and Veit-Haibach, P. (2020). Convolutional neural networks for improving image quality with noisy pet data. *EJNMMI research*, 10:1–11.

[65] Serrano-Sosa, M., Spuhler, K., DeLorenzo, C., and Huang, C. (2020). Denoising low-count pet images using a dilated convolutional neural network for kinetic modeling.

[66] Shao, R. and Bi, X.-J. (2022). Transformers meet small datasets. *IEEE Access*, 10:118454–118464.

[67] Sikka, A., Virk, J. S., Bathula, D. R., et al. (2021). Mri to pet cross-modality translation using globally and locally aware gan (gla-gan) for multi-modal diagnosis of alzheimer's disease. *arXiv preprint arXiv:2108.02160*.

[68] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

[69] Spuhler, K., Serrano-Sosa, M., Cattell, R., DeLorenzo, C., and Huang, C. (2020). Full-count pet recovery from low-count image using a dilated convolutional neural network. *Medical Physics*, 47(10):4928–4938.

[70] Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852.

[71] Sun, H., Jiang, Y., Yuan, J., Wang, H., Liang, D., Fan, W., Hu, Z., and Zhang, N. (2022). High-quality pet image synthesis from ultra-low-dose pet/mri using bi-task deep learning. *Quantitative Imaging in Medicine and Surgery*, 12(12):5326.

[72] Tan, B., Xue, Y., Bi, L., and Kim, J. (2024). Full-trsun: A full-resolution transformer unet for high quality pet image synthesis. In *International Workshop on Machine Learning in Medical Imaging*, pages 238–247. Springer.

[73] Tang, Z., Jiang, C., Cui, Z., and Shen, D. (2024). Hf-resdiff: High-frequency-guided residual diffusion for multi-dose pet reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 372–381. Springer.

[74] Vandenberghe, S. and Marsden, P. K. (2015). Pet-mri: a review of challenges and solutions in the development of integrated multimodality imaging. *Physics in Medicine & Biology*, 60(4):R115.

[75] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[76] Voss, S. D., Reaman, G. H., Kaste, S. C., and Slovis, T. L. (2009). The alara concept in pediatric oncology. *Pediatric radiology*, 39:1142–1146.

[77] Wang, M., Zhou, X., Jin, M., Zhang, Y., Liu, L., and Huang, G. (2024a). Multiroimix: A data augmentation method for pet/ct multimodal medical images. *Journal of Medical and Biological Engineering*, pages 1–9.

[78] Wang, Y., Luo, Y., Zu, C., Zhan, B., Jiao, Z., Wu, X., Zhou, J., Shen, D., and Zhou, L. (2024b). 3d multi-modality transformer-gan for high-quality pet reconstruction. *Medical Image Analysis*, 91:102983.

[79] Wang, Y., Yu, B., Wang, L., Zu, C., Lalush, D. S., Lin, W., Wu, X., Zhou, J., Shen, D., and Zhou, L. (2018). 3d conditional generative adversarial networks for high-quality pet image estimation at low dose. *Neuroimage*, 174:550–562.

[80] Wang, Y.-R., Baratto, L., Hawk, K. E., Theruvath, A. J., Pribnow, A., Thakor, A. S., Gatidis, S., Lu, R., Gummidipundi, S. E., Garcia-Diaz, J., et al. (2021). Artificial intelligence enables whole-body positron emission tomography scans with minimal radiation exposure. *European journal of nuclear medicine and molecular imaging*, 48:2771–2781.

[81] Wiatrak, M., Albrecht, S. V., and Nystrom, A. (2019). Stabilizing generative adversarial networks: A survey. *arXiv preprint arXiv:1910.00927*.

[82] Witney, T. H. and Lewis, D. Y. (2019). Imaging cancer metabolism with positron emission tomography (pet). *Cancer Metabolism: Methods and Protocols*, pages 29–44.

[83] Xiang, L., Qiao, Y., Nie, D., An, L., Lin, W., Wang, Q., and Shen, D. (2017). Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose pet/mri. *Neurocomputing*, 267:406–416.

[84] Xue, H., Teng, Y., Tie, C., Wan, Q., Wu, J., Li, M., Liang, G., Liang, D., Liu, X., Zheng, H., et al. (2020). A 3d attention residual encoder–decoder least-square gan for low-count pet denoising. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 983:164638.

[85] Xue, Y., Bi, L., Peng, Y., Fulham, M., Feng, D. D., and Kim, J. (2023a). Pet synthesis via self-supervised adaptive residual estimation generative adversarial network. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 8(4):426–438.

[86] Xue, Y., Peng, Y., Bi, L., Feng, D., and Kim, J. (2023b). Cg-3dsrgan: A classification guided 3d generative adversarial network for image quality recovery from low-dose pet images. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 1–4. IEEE.

[87] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4):1–39.

[88] Yang, Z., Chen, H., Qian, Z., Zhou, Y., Zhang, H., Zhao, D., Wei, B., and Xu, Y. (2024). Region attention transformer for medical image restoration. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–613. Springer.

[89] Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., and Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032.

[90] Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., and Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739.

[91] Zeng, P., Zhou, L., Zu, C., Zeng, X., Jiao, Z., Wu, X., Zhou, J., Shen, D., and Wang, Y. (2022). 3d cvt-gan: a 3d convolutional vision transformer-gan for pet reconstruction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 516–526. Springer.

[92] Zhang, K., Zuo, W., Chen, Y., Meng, D., and Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155.

[93] Zhang, L., Xiao, Z., Zhou, C., Yuan, J., He, Q., Yang, Y., Liu, X., Liang, D., Zheng, H., Fan, W., et al. (2022). Spatial adaptive and transformer fusion network (stfnet) for low-count pet blind denoising with mri. *Medical Physics*, 49(1):343–356.

[94] Zhang, X., Liu, C., Ou, N., Zeng, X., Zhuo, Z., Duan, Y., Xiong, X., Yu, Y., Liu, Z., Liu, Y., et al. (2023). Carvemix: a simple data augmentation method for brain lesion segmentation. *Neuroimage*, 271:120041.

[95] Zhao, K., Zhou, L., Gao, S., Wang, X., Wang, Y., Zhao, X., Wang, H., Liu, K., Zhu, Y., and Ye, H. (2020). Study of low-dose pet image recovery using supervised learning with cyclegan. *Plos one*, 15(9):e0238455.

[96] Zhou, L., Schaefferkoetter, J. D., Tham, I. W., Huang, G., and Yan, J. (2020). Supervised learning with cyclegan for low-dose fdg pet image denoising. *Medical image analysis*, 65:101770.

[97] Zuo, S., Xiao, Y., Chang, X., and Wang, X. (2022). Vision transformers for dense prediction: A survey. *Knowledge-Based Systems*, 253:109552.