THE UNIVERSITY OF SYDNEY

MASTER THESIS

Diffusion Model-Based Reconstruction of Low-Dose PET to Standard-Dose PET

Author:
Qingcheng LYU

Supervisor:
A/Prof. Luping ZHOU

Co-Supervisor:
A/Prof. Dong YUAN

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

in the

School of Electrical and Computer Engineering
Faculty of Engineering

2025

Abstract of thesis entitled

Diffusion Model-Based Reconstruction of Low-Dose PET to Standard-Dose PET

Submitted by

Qingcheng LYU

for the degree of Master of Philosophy

at The University of Sydney

in June, 2025

Positron Emission Tomography (PET) is a powerful functional imaging modality widely used in clinical diagnostics and biomedical research, owing to its unique ability to visualize metabolic processes in vivo. However, due to inherent limitations in the image acquisition process, achieving high-quality PET images in clinical practices typically requires administering higher doses of radioactive tracers, which can introduce potential health risks to patients. Conversely, reducing the injected tracer dosage can mitigate these risks, but resulting in images that are significantly noisier and less clear, commonly referred to as low-dose PET (LPET) images. Given the critical role of PET in modern medicine, there is a growing research interest in developing advanced reconstruction techniques that can effectively reconstruct LPET images into high-quality standard-dose PET (SPET) images. The challenge of accurately and efficiently reconstructing high-quality PET images from low-dose, or even ultra-low-dose (UDPET), acquisitions is of paramount importance, as it holds the promise of enhancing diagnostic accuracy while minimizing patient exposure to radiation.

In this thesis, we begin by presenting an overview of the problem setting, detailing the inherent challenges and the motivations underpinning our research, thereby establishing a solid foundation for understanding the task at hand. Next, we provide a comprehensive literature review that spans PET imaging, conventional medical image reconstruction methods, deep learning-based reconstruction techniques, the theoretical underpinnings of diffusion models, and their application in medical image reconstruction. Based on the insights drawn from this review, we propose a novel wavelet-informed diffusion **WiD-PET** method designed to address critical limitations of existing approaches—namely, low computational efficiency, suboptimal detail restoration, and spatial discontinuity. Building on this foundation, we further propose **CWD-PET**. Finally, we discuss potential research directions and outline opportunities for future refinement in the domain of low-dose PET (LPET) reconstruction using diffusion models.

Keywords: Positron Emission Tomography; Diffusion Model; Image Reconstruction; Ultra-low Doses PET; Wavelet Transformation; Cross-dose Reconstruction; Multi-Modal Learning.

Diffusion Model-Based Reconstruction of Low-Dose PET to Standard-Dose PET

by

Qingcheng LYU
B.E. *Beijing Institute of Technology*

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy

at

University of Sydney
June, 2025

Declaration

I, Qingcheng LYU, declare that this thesis titled, "Diffusion Model-Based Reconstruction of Low-Dose PET to Standard-Dose PET", which is submitted in fulfillment of the requirements for the Degree of Master of Philosophy, represents my own work except where due acknowledgment has been made. I further declare that it has not been previously included in a thesis, dissertation, or report submitted to this University or to any other institution for a degree, diploma or other qualifications.

Signed: _____

Date: June 10, 2025

For My Loving Family and Friends

Acknowledgements

With my greatest respect, I would like to express my deepest gratitude to all those who have supported me throughout the process of my Master of Philosophy.

First, I want to most sincerely thank to my supervisors, A/Prof. Luping Zhou and A/Prof. Dong Yuan, whose invaluable guidance and constant encouragement have been instrumental in this meaningful journey. I especially wish to thank A/Prof. Zhou, whose dedicated mentorship has been invaluable. Her guidance not only helped me acquire knowledge but also provide me with profound insights that have shaped my academic journey.

Then, I would like to sincerely thank my family. They are my foundation, my driving power, and my guiding light. Despite being separated by thousands of miles, they are standing by my shoulders, propelling me forward.

Next, I would like to sincerely thank my colleagues, classmates and peers. Throughout this journey, their academic support and research assistance have always been indispensable.

Lastly, I wish to show my heartfelt gratitude to all the friends who supported me. Your companionship and insights from different fields helped me navigate challenges and celebrate success. I am especially grateful to Joe. G, and XR. L, and most importantly Bitan Song, for being a constant source of encouragement and insight.

Once again, thanks to all the people who cared for me and supported me, I wouldn't have been able to go this far without your company, thank you!

Qingcheng LYU
University of Sydney
June 10, 2025

Attribution Statement

I, Qingcheng LYU, hereby acknowledge that while the research work presented in this thesis is entirely my own, I have utilized generative artificial intelligence tools (e.g., ChatGPT) to assist in refining and optimizing the written content. All suggestions and outputs from these tools were carefully reviewed, edited, and integrated by me, ensuring that the final document accurately reflects my original ideas and research contributions. Any external sources or assistance have been duly acknowledged throughout the thesis.

Chapter 4 of this thesis has been submitted to MICCAI 2025, and Chapter 5 of this thesis is to be submitted to TMI. I designed the study, analyzed the data, and wrote the manuscript. In addition to the authorship attribution 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.

Student Signed: _____

Date: June 10, 2025

Supervisor Signed: _____

Date: June 10, 2025

Supervisor Confirmation

As supervisor of the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor: Luping Zhou

Signed: _____

Date: _____

List of Publications

Journals:

[1] **Qingcheng Lyu**, and Luping Zhou, "CWD-PET: Cross-dose PET image reconstruction by Wavelet-informed Diffusion model with Fast Inference", *IEEE Transactions on Medical Imaging (TMI)*. (To be submitted to TMI)

Conferences:

[1] **Qingcheng Lyu**, Tong Chen, Erjian Guo, Yiran Wang, and Luping Zhou, "WiD-PET: PET Image Reconstruction from Low-Dose Data Using a Wavelet-informed Diffusion Model with Fast Inference", in *The Medical Image Computing and Computer Assisted Intervention (MICCAI) 2025*. (Under review)

Contents

List of Figures

List of Tables

List of Abbreviations

cDDPM	**c**onditional **D**iffusion **D**enoising **I**mplicit **M**odel
CLIP	**C**ontrastive **L**anguage-**I**mage **P**re-Training
CNN	**C**onvolutional **N**eural **N**etwork
CoC	**Co**ntext **C**lusters
CPM	**C**oarse **P**rediction **M**odule
CT	**C**omputerized **T**omography
CWD	**C**ross-dose **W**avelet-informed **D**iffusion
DDIM	**D**iffusion **D**enoising **I**mplicit **M**odel
DDPM	**D**iffusion **D**enoising **P**robabilistic **M**odel
DWT	**D**ecompose **W**avelet **T**ransformation
FDG	**F**luoro**d**eoxy**g**lucose
GAN	**G**enerative **A**dversarial **N**etwork
GPU	**G**raphics **P**rocessing **U**nit
HFE	**H**igh **F**requency **E**nhancer
IDWT	**I**nverse **D**ecompose **W**avelet **T**ransformation
IRM	**I**terative **R**efinement **M**odule
LPET	**L**ow-dose **P**ositron **E**mission **T**omography
MBq	**M**ega**B**ec**q**uerels
mCi	**m**illi**c**uri**e**s
ML	**M**aximum **L**ikelihood
MRI	**M**agnetic **R**esonance **I**maging
M^3**TRec**	**M**ulti-**m**odal **M**asked **T**ext **Rec**onstruction
NMSE	**N**ormalized **M**ean **S**quared **E**rror
OMTA	**Optimal** **M**ulti-modal **T**ransport co-**A**ttention
PEF	**P**rompt **E**mbedding **F**using
PET	**P**ositron **E**mission **T**omography
PSNR	**P**eak **S**ignal-to-**N**oise **R**atio
RFN	**R**e**f**inement-**N**et

SCA	**S**patial **C**onsistency **A**ttention
SCFE	**S**patial **C**onsistency **F**eature **E**xtracter
SNR	**S**ignal-to-**N**oise **R**atio
SOTA	**S**tate-**O**f-**T**he-**A**rt
SPET	**S**tandard-dose **P**ositron **E**mission **T**omography
SR	**S**parse **R**epresentation
SSIM	**S**tructural **S**imilarity **I**ndex **M**easure
SSTDL	**S**emi-**s**upervised **T**ripled **D**ictionary **L**earning
TCoC	**T**ransposed **C**omputerized **T**omography
UDPET	**U**ltra-low **Dose** **P**ositron **E**mission **T**omography
WID	**W**avelet **I**nforomed **D**iffusion

Chapter 1

Introduction

In this chapter, we begin by outlining the critical problem of denoising and reconstructing PET medical images. We then detail the primary challenges encountered in current PET image denoising and reconstruction methods, thereby clarifying the motivations behind our research. Finally, we provide an overview of the thesis structure and summarize our contributions in addressing these challenges through diffusion-based approaches.

1.1 Problem Statement

Positron Emission Tomography (PET) is a critical imaging modality that plays a pivotal role in clinical diagnostics by providing insights into in vivo metabolic processes. The clinical efficacy of PET imaging is heavily dependent on image quality, as high-quality PET images are essential for accurate diagnosis and effective treatment planning [5, 16, 45, 48, 60]. However, acquiring such high-quality images typically requires administering high doses of radioactive tracers, which may cause a potential [6, 36, 42] risk of radiation exposure to patients. In contrast, low-dose PET (LPET) imaging—though safer—results in images with elevated noise levels and reduced clarity. Therefore, it becomes a crucial problem for the acquisition of high-quality SPET images from LPET images.

Consequently, many methods have been proposed to solve the problem. This reconstruction task is fundamentally a denoising problem,

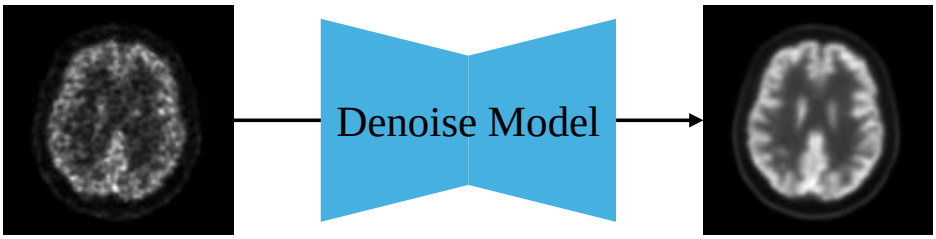


Figure 1.1: An example of the PET reconstruction process. The input and output correspond to 1/100 (low-dose PET) and 1/1 (standard-dose PET) dose levels from the dataset, with the blurriness reflecting the intrinsic characteristics of PET imaging.

wherein the goal is to recover essential image details from noisy, low-information inputs. Despite numerous advances, existing approaches still face significant challenges: they often exhibit suboptimal performance, and inadequate handling of ultra-low-dose PET (UDPET) scenarios, where the tracer dose may be as low as 1/20, 1/50, or even 1/100 of the standard dose. Moreover, many methods are restricted to processing a single dose level, further limiting their clinical applicability. These limitations can lead to poor image quality in the reconstructed PET images, thereby potentially compromising the accuracy of subsequent clinical diagnoses. Thus, developing an efficient and high-quality reconstruction method is of paramount importance.

Figure 1.1 illustrates a simplified overview of the PET image reconstruction process. Typically, an LPET image is obtained alongside its corresponding SPET image. The objective is to design a reconstruction model that can transform an LPET image into one that matches the quality of an SPET image. Conventional reconstruction frameworks generally incorporate components such as feature extractors and up/down-sampling modules. The critical challenge, however, lies in extracting and restoring sufficient information from the inherently noisy LPET images—especially in cases of ultra-low-dose imaging—where the available information is extremely limited. Traditional methods have been

employed to address this problem, yet their performance remains inadequate for the most challenging low-dose conditions. Recent advances in convolutional neural networks (CNNs) and deep learning methods have introduced more sophisticated approaches, and the advent of diffusion models, with their inherent denoising capabilities, offers a promising direction for overcoming these challenges.

1.2 Challenges and Motivations

As discussed in the previous section, reconstructing high-quality standard-dose PET (SPET) images from low-dose PET (LPET) images poses numerous challenges, particularly due to the distinct characteristics inherent in various foundational approaches.

Reconstructing high-quality standard-dose PET (SPET) images from low-dose PET (LPET) images poses numerous challenges, largely due to the inherent characteristics of the imaging process and the limitations of existing reconstruction approaches. Foremost among these challenges is the task of effectively recovering high-quality SPET images from their corresponding LPET counterparts. In PET imaging, the administered dose of radioactive tracers directly influences image quality. Radiotracers are designed to accumulate preferentially in regions exhibiting high metabolic activity, enabling the scanner to capture detailed physiological information. However, when a lower dose is used—especially in ultra-low-dose scenarios—the resulting LPET images are marred by increased noise and artifacts [6, 36, 42]. Moreover, key structural details such as edges and textures become blurred and degraded, further complicating the reconstruction process. As the dose decreases, the information available for effective reconstruction diminishes, placing additional strain on the learning capabilities of reconstruction models. These inherent limitations introduce significant obstacles to achieving high-quality SPET reconstructions, with the challenges becoming even more pronounced under ultra-low-dose conditions. More details regarding these challenges will be discussed in Chapter 2.

Apart from these image quality issues, some methods are driven from deep learning models, including those employing Generative Adversarial Networks (GANs) and, more notably, diffusion-based models—face a significant computational cost and inference time cost challenge. Deep learning architectures typically rely on deeper networks and more complex structures. When applied to full-resolution LPET reconstruction tasks, these models incur substantial computational demands, particularly for high-resolution images. Diffusion-based models, in particular, require multiple iterative denoising steps to progressively refine the reconstructed image. This iterative process involves repeated passes through deep neural networks, leading to extended inference times—a well-recognized limitation in standard diffusion frameworks, and one that is even more acute in the context of PET image denoising and reconstruction. Consequently, reducing computational cost and shortening inference time remain critical challenges to achieving practical and efficient reconstruction outcomes.

Another significant challenge is the spatial discontinuity problem. Existing 2D GAN-based and diffusion-based methods do not directly process 3D LPET images; rather, they operate on individual 2D slices, which are subsequently reassembled into a complete 3D volume for further analysis or clinical application. Since the model reconstructs each slice independently, inconsistencies can arise between adjacent slices, leading to spatial discontinuities in the final 3D image. These discontinuities can severely compromise overall reconstruction quality and, by extension, the reliability of clinical diagnoses. Addressing this issue is therefore of paramount importance.

To address the above three challenges, we propose a novel framework termed the Wavelet-Informed Diffusion (WiD-PET) framework. The WiD-PET framework incorporates a 2D-wavelet-Informed diffusion strategy, the goal of this framework is to accelerate inference speed, and it is coupled with a specially designed High-Frequency Enhancer (HFE) module that facilitates the reconstruction of high-quality SPET images with well-preserved details and textures. In addition, to specifically tackle the spatial discontinuity problem, we introduce a novel Spatial

Consistency Feature Extractor (SCFE) along with a Spatial Consistency Attention (SCA) module. The combination of these components effectively mitigates inconsistencies across 2D slices, thereby enhancing the continuity of the reconstructed 3D volume.

Last but not least, existing approaches primarily focus on optimizing reconstruction performance for a single-dose level, overlooking the challenge of cross-multi-dose reconstruction. In clinical practice, the acquisition protocols for LPET images vary considerably, meaning that a model tailored to a single dose level may not satisfy the diverse requirements encountered in real-world applications. It is therefore imperative to develop a framework capable of handling LPET images acquired at different dose levels while maintaining high reconstruction quality in terms of fidelity, detail preservation, texture accuracy, and inference speed. To address this fourth challenge, we propose a novel Cross-dose Refine Wavelet-informed Diffusion (CWD-PET) framework that builds upon the WID backbone and incorporates the HFE, SCFE, and SCA modules. The innovative aspect of this framework is the inclusion of a Prompt-Embedding Fusing (PEF) module, which integrates externally provided prompt information with time embeddings. Additionally, we introduce a novel lightweight Refinement-Net (RFN) that further enhances the model's ability to cope with varying dose levels. Together, these design innovations enable robust performance across multiple dose levels in PET image reconstruction.

In summary, this thesis focuses on developing novel LPET reconstruction methods based on diffusion models to address four main challenges: (1) reconstructing high-quality SPET images from inherently noisy LPET data, (2) reducing the computational cost and inference time, (3) overcoming spatial discontinuities arising from 2D slice-based reconstruction approaches, and (4) enabling robust multi-dose level reconstruction. Through the proposed methodologies and innovations, we aim to significantly advance the state-of-the-art in PET image denoising and reconstruction, thereby enhancing clinical diagnostic accuracy and efficiency.

1.3 Thesis Outline and Contributions

The remainder of this thesis is organized into five chapters, each addressing different aspects of the research:

Chapter 2. Literature Review. This chapter presents a comprehensive review of the development of PET medical image reconstruction. It covers foundational knowledge of PET imaging and discusses related works pertinent to the two methods proposed in this thesis, offering a rich discussion of existing approaches and their limitations.

Chapter 3. Background Introduction of Low Dose PET Reconstruction using Diffusion Technology. In this chapter, we delve further into the problem statement and introduce the current evaluation methodologies and datasets used in PET medical image reconstruction. This background provides the necessary context for understanding the challenges and the performance metrics relevant to our research.

Chapter 4. Standard-Dose PET Reconstruction from Low-Dose PET by Wavelet-Informed Diffusion Model with Fast Inference. Here, we propose a fast wavelet-Informed diffusion model designed to reconstruct high-quality SPET images from UDPET images while preserving fine details and textures. We compare our results with state-of-the-art competitors across three different ultra-low-dose levels, demonstrating the superior performance of our approach.

- The contributions in this part are included in:

> **Qingcheng Lyu**, Tong Chen, Erjian Guo, Yiran Wang, and Luping Zhou, "WiD-PET: PET Image Reconstruction from Low-Dose Data Using a Wavelet-informed Diffusion Model with Fast Inference", *The Medical Image Computing and Computer Assisted Intervention (MICCAI) 2025*. (Under review)

Chapter 5. Standard-Dose PET Reconstruction from Multi-dose LPET by Cross-dose Wavelet-informed Refine Diffusion Model This chapter introduces a novel framework capable of reconstructing SPET images from LPET images of multiple dose levels using a single model.

By employing a specifically designed prompt-embedding fusing module (PEF) along with a Refinement-Net (RFN), this method outperforms current state-of-the-art competitors at three different ultra-low-dose levels, showcasing remarkable overall performance.

- The contributions in this part are included in:

> **Qingcheng Lyu**, and Luping Zhou, "CWD-PET: Cross-dose PET image reconstruction by Wavelet-Informed Diffusion model with Fast Inference", *IEEE Transactions on Medical Imaging*. (To be submitted to TMI)

Chapter 6. Conclusion and Future Work. In the final chapter, we present our conclusions based on the research findings and contributions of this thesis. We also discuss potential directions for future research, highlighting possible refinements and advancements in the field.

Chapter 2

Literature Review

In this chapter, we begin by presenting the background of PET medical imaging, outlining its fundamental principles and clinical significance. We then review traditional PET image reconstruction methods, followed by an examination of deep learning-based reconstruction techniques. Finally, we introduce diffusion denoising probabilistic models (DDPM) and discuss related works in PET denoising tasks.

2.1 PET Medical Imaging

It is well known that medical imaging constitutes one of the most essential components of modern nuclear medicine procedures and clinical applications, as it provides a robust tool for obtaining multidimensional information crucial for diagnosis, prognosis, and therapy response monitoring [19, 27, 61]. Among the various modalities, positron emission tomography (PET) has emerged as a mature and powerful technique for visualizing metabolic processes within the human body [5, 16, 48, 60]. PET imaging relies on the detection of positron-emitting radiotracers—chemical compounds labeled with isotopes such as 18F or 11C [20, 38, 57]. In a typical PET examination, the radiotracer is administered intravenously. Once injected, the radiotracer distributes throughout the body according to its biochemical properties and preferentially accumulates in regions exhibiting high metabolic activity. As the isotope decays, positrons are emitted; these positrons rapidly annihilate with nearby electrons, producing pairs of gamma photons that travel in nearly opposite directions. The PET scanner, equipped with an array of scintillator

detectors arranged in a ring-shaped configuration around the patient, captures these photons. The collected data is subsequently used to reconstruct a three-dimensional image of the tracer distribution, providing critical insights into both physiological and pathological processes. PET imaging is indispensable for diagnosing internal organs and is widely applied in the evaluation of brain function, lung pathology, tumor detection, and various cancer diagnoses.

Given the pivotal role of PET in clinical applications [3, 34], precise dose administration is fundamental to optimizing imaging performance. The dosage is typically measured with the unit in megabecquerels (MBq) or millicuries (mCi) [52], and must be carefully calibrated to balance image quality with patient safety. A higher injected dose enhances the signal-to-noise ratio (SNR) and improves spatial resolution, yielding clearer and more detailed images with reduced noise. However, this advantage comes at the cost of increased radiation exposure, which can be particularly concerning for repeated scans or vulnerable patient populations. Conversely, a lower dose minimizes radiation risk but often results in images with elevated noise levels and diminished diagnostic detail and texture. This critical trade-off between dose level and image quality has spurred extensive research into advanced reconstruction (or denoising) algorithms capable of effectively mitigating noise and enhancing image fidelity under low-dose conditions. Such innovations are essential to ensure that PET imaging remains both diagnostically powerful and safe for clinical use [1, 24, 31, 37, 50, 51, 65].

Over the past few years, numerous researchers have dedicated their efforts to developing traditional PET image reconstruction methods. With the advent of deep learning and modern image reconstruction techniques, even more advanced approaches have emerged. These developments will be discussed in detail in the following section.

2.2 Conventional PET Reconstruction Methods

2.2.1 Overview

Traditional methods for PET image reconstruction have been extensively studied and refined over the past decades, particularly during the period from 2010 to 2016 [9, 43, 53, 68]. These techniques are primarily based on mathematical and statistical models designed to solve the inverse problem inherent in reconstructing high-quality standard-dose PET (SPET) images from low-dose PET (LPET) data. Early approaches, such as analytical methods and iterative algorithms, laid the foundation by using well-established principles to manage issues like noise, limited resolution, and the ill-posed nature of the reconstruction task.

Due to inherent performance limitations—especially when relying solely on the information provided by LPET images—researchers have sought to augment these methods by integrating data from additional imaging modalities, such as MRI or CT [41, 68]. The rationale behind this multi-modality approach is to supplement the limited and noisy LPET data with richer, complementary anatomical or functional information, thereby enhancing the reconstruction quality.

This overview sets the stage for a deeper discussion of specific traditional reconstruction methods. Subsequent sections will review representative approaches, such as patch-based sparse representation methods, kernel-based reconstruction techniques, and regression forest frameworks, each of which has contributed uniquely to advancing PET image reconstruction. These methods, while effective to varying degrees, also highlight the challenges and trade-offs that continue to motivate the development of more advanced techniques, including those based on deep learning and diffusion models.

2.2.2 Key Techniques in Conventional Reconstruction

In their work, Yan Wang et al. proposed a framework based on patch-based sparse representation (SR) methods [67], which utilizes complete

paired datasets consisting of LPET, corresponding SPET, and MRI images. To overcome the challenge posed by the unavailability of paired multi-modality data in some cases—a significant drawback of SR-based methods—they introduced a semi-supervised tripled dictionary learning (SSTDL) approach. This strategy allows the model to be trained on incomplete samples while still achieving effective reconstruction.

Similarly, Guobao Wang et al. proposed an LPET reconstruction method based on kernel methods [66]. In their approach, the intensity of each pixel in the LPET image is modeled as a function of a set of features extracted from prior information. This model is then integrated into the forward model of PET projection data, and the coefficients are estimated using Maximum Likelihood (ML) or penalized likelihood reconstruction methods. Compared with traditional reconstruction techniques without post-reconstruction denoising, their approach offers improved bias characteristics and enhanced contrast recovery.

Additionally, Jiayin Tang et al. developed a reconstruction method based on regression forests [41], which also incorporates corresponding MRI images alongside LPET and SPET pairs to provide additional information. Their framework extracts features from different patches, corresponding to various organs in the MRI images, to build tissue-specific models. These models are first used to generate an initial prediction of the SPET image, and an iterative refinement process is subsequently applied to achieve superior reconstruction results.

2.2.3 Challenges and Limitations of Conventional Methods

Despite their contributions, traditional reconstruction methods face several limitations. First, the reliance on additional modalities such as MRI or CT to supplement LPET data introduces complexity and dependency on the availability of multi-modal datasets. Second, methods based on sparse representation or dictionary learning, while effective in leveraging spatial redundancies, often struggle with incomplete or inconsistent

data, which can adversely affect reconstruction quality. Third, kernel-based approaches and regression forest methods may achieve improvements in bias and contrast recovery, but they typically require careful feature selection and parameter tuning, which limits their robustness and generalizability across different dose levels and imaging conditions.

Traditional image denoising techniques, including classical filtering methods, wavelet thresholding, and non-local means, often rely on simplified noise models and local image statistics, which limit their effectiveness in addressing the complex and spatially varying noise characteristics inherent in low-dose PET data [7, 15]. Furthermore, these methods generally treat denoising as a separate pre-processing step, without fully exploiting the structural and contextual information available across multiple image slices or modalities. As a result, their denoising performance on ultra-low-dose PET images—where noise is more severe and signal degradation is significant—is often suboptimal. Traditional denoising methods consequently struggle to recover clinically relevant details, making them inadequate for high-quality PET reconstruction tasks. Therefore, modern PET reconstruction research has largely moved beyond these conventional denoising approaches in favor of more integrated and data-driven methods.

These limitations highlight the need for more advanced reconstruction techniques that can robustly and efficiently handle the challenges inherent in low-dose PET imaging without heavy reliance on additional modalities. Subsequent sections will explore the evolution of reconstruction methods from these traditional approaches to more recent deep learning and diffusion-based techniques.

2.3 Deep Learning-Based PET Reconstruction

2.3.1 Overview

With the development of deep learning techniques, the field of medical image reconstruction has been revolutionized. Leveraging large-scale

datasets, deep learning models are capable of learning complex mappings between input and output images, thereby significantly enhancing image quality compared to traditional methods. Unlike conventional approaches—which often rely on hand-crafted features and relatively shallow models—deep learning methods automatically extract hierarchical features through deep network architectures. This allows them to capture both low-level textures and high-level semantic information, leading to more accurate and robust reconstructions [12, 55].

Convolutional Neural Networks (CNNs) serve as the cornerstone of many deep learning approaches in medical imaging [35, 58], owing to their strong feature extraction capabilities and ability to capture intricate spatial details [44]. Building upon the basic CNN architecture, specialized structures such as U-Net have been developed specifically for image reconstruction tasks. U-Net employs an encoder-decoder architecture with skip connections that help recover fine-grained details lost during down-sampling, thereby yielding superior reconstruction performance [8, 21].

Moreover, advanced frameworks like Generative Adversarial Networks (GANs) [25] extend the capabilities of CNNs by incorporating an adversarial training paradigm. This approach encourages the generation of images that closely resemble high-quality target images, effectively pushing the network to produce more realistic reconstructions. In addition to these, the concept of residual learning—as exemplified by ResNet [30]—introduces residual blocks that mitigate the vanishing gradient problem, enabling the training of very deep networks. These ResNet modules, which can be integrated into various architectures, enhance feature learning and overall reconstruction performance. Meanwhile, autoencoders [69] facilitate effective unsupervised learning by compressing input data into a latent representation and then reconstructing the image, proving especially valuable for denoising and dimensionality reduction tasks.

Collectively, these architectures—CNNs, U-Net, GANs, ResNet, and autoencoders—form the foundational building blocks of many advanced

frameworks in medical image reconstruction. Their complementary design principles and inherent strengths have made them indispensable tools, continually driving progress and innovation in the field.

2.3.2 Applications in PET Reconstruction

Several recent studies have demonstrated the potential of deep learning techniques for PET image reconstruction by leveraging novel network architectures and innovative training strategies. These methods not only improve image quality but also address challenges such as noise reduction and detail preservation in low-dose scenarios, thereby enhancing diagnostic accuracy in clinical applications.

In one notable work, Kuang et al. [22] combined the strengths of CNNs with residual connections to propose a deep residual CNN framework designed to extract patient-specific information and enhance the quality of PET images. In their approach, low-dose PET images are used as inputs while the corresponding high-dose images serve as labels, enabling the model to learn a direct mapping between degraded and high-quality images. By learning this mapping, their framework effectively reconstructs higher-quality PET images from the degraded inputs. Additionally, they introduced an iterative training strategy aimed at progressively refining the reconstruction quality, which further improves the performance and robustness of their model. This iterative refinement allows the network to focus on subtle image details in subsequent passes, thereby reducing artifacts and enhancing overall image fidelity.

In another study, Tan et al. proposed a standard-dose PET (SPET) reconstruction framework based on the U-Net architecture [64]. Recognizing the importance of preserving fine-grained details and capturing long-range dependencies, Tan et al. incorporated a transformer block into the U-Net backbone. This hybrid design allowed the model to benefit from the robust spatial feature extraction of CNNs in the encoder-decoder structure, while the transformer block contributed to modeling global contextual relationships. As a result, the combined architecture was able to address both local and global aspects of the image, ultimately

yielding improved reconstruction outcomes. Their work underscores the importance of integrating attention mechanisms to capture context that might be missed by conventional convolutional operations alone.

Moreover, Luo et al. introduced a GAN-based framework that similarly integrated transformer blocks into its design [46]. Their architecture—dubbed the EncoderCNN-Transformer-DecoderCNN, which capitalizes on the complementary strengths of transformers and CNNs. While the transformer components excel at capturing global features and long-range semantic information, the CNN modules efficiently extract rich spatial features with a compact representation. This synergistic combination enables their GAN-based model to produce high-fidelity SPET reconstructions from low-dose PET (LPET) inputs. The adversarial training framework further refines the output by encouraging the generator to produce images that are not only quantitatively accurate but also perceptually realistic, which is critical for clinical acceptance.

In addition to these approaches, Cui et al. proposed an innovative 3D point-based context clusters GAN [13]. Diverging from conventional GAN architectures, this method generates a predicted image based on a point-cluster strategy. The process begins with transforming the input LPET image into point clusters, which serve as a compact representation of the image's spatial structure. Subsequently, the network employs four Context Clusters (CoC) blocks followed by four Transposed Context Clusters (TCoC) blocks to predict the residual points between the LPET and the desired SPET images. These predicted residuals are then added back to the original point clusters, effectively generating a refined SPET image prediction. This novel strategy enables the model to capture and preserve subtle structural details that are often lost in traditional reconstruction approaches, thereby improving both the visual quality and clinical utility of the reconstructed images.

Collectively, these studies illustrate the diverse strategies adopted in deep learning-based PET reconstruction. By integrating CNNs, residual learning, transformer blocks, and GAN frameworks, these methods demonstrate significant improvements in image quality and robustness. The innovative use of hybrid architectures and iterative refinement

strategies offers promising avenues for further research and clinical application in PET imaging, paving the way for models that can reliably reconstruct high-quality images even under challenging low-dose conditions.

2.3.3 Challenges and Limitations of Deep Learning-Based Methods

Despite their significant advancements, deep learning methods for PET reconstruction still exhibit several notable limitations. CNN-based approaches, for instance, rely heavily on convolutional operations to extract local features, which works well for capturing fine spatial details. However, these methods may struggle with modeling long-range dependencies inherent in complex PET images. As a result, while CNN-based networks often yield good local reconstruction performance, their effectiveness can diminish when reconstructing global image consistency, especially when subtle variations and long-distance contextual information are critical [2].

Building on the basic CNN architecture, GAN-based methods have been widely regarded as a promising solution for PET reconstruction due to their ability to generate more realistic images through adversarial training. GANs have indeed contributed impressive results in many studies by pushing the reconstructed images closer to high-quality targets. Nevertheless, these networks are also prone to well-known issues such as mode collapse, where the generator may converge to a limited set of outputs, thereby reducing the diversity and fidelity of the reconstructions. Furthermore, despite their potential for high-quality image synthesis, GAN-based methods sometimes produce results that are inconsistent or fail to capture certain diagnostic details, leading to suboptimal image restoration performance in some scenarios [71].

In summary, while deep learning methods—including both CNN-based and GAN-based frameworks—have advanced the state-of-the-art in PET reconstruction, they still face challenges in balancing local and

global feature extraction and ensuring stable and consistent image generation. These limitations highlight the need for further innovation, paving the way for exploring alternative or hybrid approaches, such as diffusion models, which may offer improved robustness and reconstruction fidelity in low-dose PET imaging.

2.4 Diffusion-Based Methods for PET Reconstruction

2.4.1 Overview

In recent years, diffusion models have emerged as a powerful class of generative models, garnering significant attention for their ability to generate high-quality images and capture complex data distributions. Their underlying concept is inspired by diffusion phenomena observed in thermodynamics [62]. At a high level, diffusion-based models work by gradually corrupting a data sample (usually is the high-quality target image) with noise over a series of steps. By the end of this forward noising process, the original data distribution is essentially masked by a noisy distribution, resulting in the loss of the original information. The key idea, however, is to learn a reverse process that can progressively remove the added noise, ultimately reconstructing the original data from a pure noise input. This two-phase process—comprising a forward noising phase and a reverse denoising phase—forms the backbone of a typical diffusion model.

A seminal work in this area was proposed by Ho et al., who introduced Denoising Diffusion Probabilistic Models (DDPM) [33]. Unlike traditional CNN or GAN architectures, DDPM leverages a probabilistic framework to model the diffusion process, which not only leads to more stable training but also produces images with remarkably high fidelity. The DDPM framework operates by learning to predict either the added noise or directly the denoised image at each step, thus gradually reversing the corruption process. The inherent advantages of DDPM include

its capacity for capturing fine details and preserving global structures, which are critical for high-quality image reconstruction.

Conditional Denoising Diffusion Probabilistic Models (cDDPM) extend the original DDPM framework by incorporating additional conditioning information into the diffusion process. This conditioning can take various forms, such as class labels, segmentation masks, or other auxiliary data, and is used to guide the generation process toward desired outputs. In the context of image reconstruction, particularly in medical imaging, cDDPM enables the model to leverage contextual or prior information (e.g., low-dose PET images, corresponding MRI data, or anatomical priors) to improve reconstruction accuracy and preserve critical diagnostic details.

These advantages have contributed to the success of DDPM in a wide range of applications—from natural image synthesis to medical image reconstruction. Owing to its robust performance and stability, numerous DDPM-based methods have been developed for medical imaging tasks. Variants have further refined the basic framework to cater to specific challenges[56, 63, 73] in various fields including reconstructing medical images. Particularly noteworthy among these advancements is the Denoising Diffusion Implicit Models (DDIM)[63]. Unlike DDPM, which employs a strictly Markovian reverse process, DDIM introduces a non-Markovian formulation that allows for a more deterministic reverse process. This deterministic approach enables DDIM to generate high-quality images with significantly fewer sampling steps. By allowing the model to "skip" intermediate steps, DDIM reduces the number of necessary denoising steps, DDIM addresses one of the primary computational challenges associated with traditional DDPM, making it an attractive alternative for real-world applications.

In summary, the theoretical soundness and practical advantages of diffusion models, exemplified by DDPM and its variants like DDIM, have positioned them as promising tools for image reconstruction. Their ability to overcome the limitations of conventional CNN and GAN-based methods has spurred extensive research, leading to the development of several specialized architectures for medical image reconstruction. These

advancements provide a solid foundation for the application of diffusion models in tasks such as PET image reconstruction, where the preservation of fine anatomical details and overall image fidelity is paramount.

2.4.2 Diffusion Models in PET reconstruction

Gong et al. proposed a framework based on a conditional DDPM (cDDPM) model for denoising SPET images from LPET inputs [23]. In their approach, after the SPET image undergoes the forward diffusion process and becomes noisy, the corresponding LPET image is used as the conditioning signal to guide the reverse denoising process. This strategy leverages the paired PET data to direct the generation process, ensuring that the restored image adheres closely to the structural and textural characteristics inherent in the LPET input. By explicitly conditioning on the LPET image, their method can effectively reduce artifacts and recover finer details, yielding promising reconstruction performance. In fact, using the LPET image as the condition has become one of the most common and effective techniques to ensure reliable and accurate denoising results in PET image reconstruction.

In another study, Jiang et al. introduced an unsupervised LPET enhancement framework (uPETe) based on a latent diffusion model [39]. Their framework features an encoder-decoder structure where the encoder maps the input SPET images into a compact latent space, and the decoder reconstructs the SPET images from the latent representation. This approach not only reduces the dimensionality of the input but also significantly decreases the computational burden during inference. To tackle the substantial inference time typically associated with diffusion models, the encoder compresses the scale of the input PET image, enabling faster processing without a major loss in image fidelity. Moreover, to keep the perturbed sample close to the actual noise distribution, they employed a Poisson diffusion process instead of the conventional Gaussian approach, which better aligns with the statistical nature of PET data. Additionally, to address the inherent information loss in LPET images, they integrated corresponding CT images to guide a cross-attention mechanism, thereby supplementing essential structural details

required for accurate reconstruction. This multi-modal integration further enhances the robustness of the reconstruction, making the approach particularly appealing in clinical settings where complementary data can be leveraged.

Cui et al. further extended the concept by exploring the use of additional information from modalities other than LPET [14]. Unlike previous methods that rely on CT or MRI images, they utilized patients' clinical tabular information to introduce multi-modal cues into the reconstruction process. Their framework incorporates a multi-modal condition encoder (Mc-Encoder) designed to extract salient features from clinical data and an optimal multi-modal transport co-attention (OMTA) module to effectively balance the input from different modalities. This design narrows the heterogeneity gap between image and tabular data, thereby mitigating semantic distortions that might otherwise occur when combining disparate data sources. Moreover, they proposed a multi-modal Masked Text Reconstruction (M^3TRec) mechanism to further fuse cross-modal inputs and enhance the overall representation. By integrating rich semantic information from non-imaging sources, this approach highlights the potential of leveraging diverse data types to guide the reconstruction process, ultimately leading to improved performance and more reliable PET image restoration.

Additionally, Han et al. proposed a hybrid framework that uses a diffusion model as the generator within a GAN architecture [28]. Their method adopts a coarse-to-fine reconstruction strategy, consisting of a coarse prediction module (CPM) followed by an iterative refinement module (IRM). In this framework, the LPET image is first processed by the CPM to produce an initial prediction that roughly captures the underlying anatomical structures. This initial output then serves as guidance for the IRM, which iteratively refines the reconstruction to enhance details and correct errors, ultimately generating a high-quality SPET prediction. Following the diffusion-based reconstruction, the generated image is further evaluated by a GAN discriminator, similar to conventional

GAN-based methods, to enforce global realism and consistency. To enrich the learning process further, the framework also incorporates auxiliary spectral guidance, which helps the model focus on capturing specific frequency components important for clinical diagnosis. Additionally, the use of negative samples during training ensures that the reconstructed images are of high quality and closely resemble the conditional LPET, effectively reducing artifacts and preserving subtle anatomical details.

In summary, these studies demonstrate a rich variety of approaches based on diffusion models for PET image reconstruction. They range from basic cDDPM frameworks that incorporate LPET as a conditioning signal to more advanced methods that leverage supplementary image data, multi-modal information, and even hybrid GAN architectures. Collectively, these methods showcase diverse strategies to overcome the challenges inherent in low-dose PET imaging. The promising performance of these approaches suggests that integrating novel architectures and additional data sources can further enhance reconstruction quality, ultimately broadening the clinical applicability of diffusion models in PET imaging.

2.4.3 Challenges and Limitations in Diffusion-Based Approaches

Despite their impressive image generation capabilities and stable training dynamics, diffusion models exhibit several notable weaknesses [4, 17, 26, 40], particularly when applied to high-resolution image reconstruction tasks.

One of the primary limitations is the inherent computational cost associated with the iterative denoising process. Diffusion models require a large number of sequential steps to gradually remove noise from an initial random input, which directly translates into prolonged inference times. This issue becomes even more pronounced for high-resolution images, as the increased data volume demands more iterations and higher computational resources to achieve a high-quality reconstruction result.

Consequently, in scenarios where rapid processing is essential—such as real-time clinical applications—the latency induced by the multiple denoising steps can become a significant bottleneck. It is worth noting that several recent approaches have attempted to address this challenge by reducing the resolution or compressing the input images to accelerate inference; however, this strategy itself introduces a trade-off and remains a common challenge when using diffusion models.

Moreover, the high computational demand of diffusion models restricts their deployment in environments with limited hardware capabilities. The need for specialized hardware, such as high-performance GPUs, to manage the iterative denoising process makes it difficult to integrate these models into routine clinical workflows or settings with constrained resources. Additionally, processing 3D images with diffusion models is particularly resource-intensive. As a result, most current research focuses on handling 2D image inputs. This focus can adversely affect 3D PET reconstruction, where the continuity between slices is crucial. Without a mechanism to adequately learn and preserve inter-slice consistency, the reconstructed 3D images may suffer from discontinuities that ultimately degrade the overall reconstruction quality (further details on this issue are discussed in Chapter 4).

Last but not least, existing diffusion-based methods predominantly address the reconstruction task for a single dose level. They are typically designed to learn the mapping from a low-dose PET (LPET) image to a standard-dose PET (SPET) image. However, in real-world clinical settings, PET imaging protocols vary, and a one-size-fits-all model may not effectively handle the diversity of dose levels encountered in practice. Thus, developing models that can robustly generalize across multiple dose scenarios remains an additional significant challenge.

In summary, while diffusion models offer significant advantages in terms of image quality and training stability, their practical application, especially for high-resolution and 3D image reconstruction, remains constrained by slow inference speeds, high computational costs, difficulties in preserving spatial continuity, and limited generalizability across different dose levels. Addressing these weaknesses is critical for enhancing

the clinical applicability of diffusion-based reconstruction methods.

2.5 Summary

In this chapter, we have provided a comprehensive overview of the key developments in PET image reconstruction, laying the groundwork for the subsequent chapters of this thesis. We began by introducing the fundamental principles of PET medical imaging, highlighting its pivotal role in clinical diagnosis and the importance of high-quality image acquisition. This was followed by an in-depth discussion of traditional reconstruction methods, which, despite their foundational contributions, suffer from inherent limitations such as noise amplification, limited resolution, and the dependency on supplementary modalities like MRI or CT to enhance reconstruction performance.

Subsequently, we explored deep learning approaches that have recently revolutionized PET image reconstruction. In particular, methods based on CNNs, U-Net, and GANs were discussed in detail. These approaches demonstrate substantial improvements in capturing complex image features and enhancing reconstruction quality. However, we also critically examined their shortcomings—including issues related to data dependency, training instability, and difficulties in preserving global image consistency—which motivate the search for alternative solutions.

The final section of our review was devoted to diffusion models, which represent a promising new direction in image reconstruction. We outlined the fundamental concepts underlying diffusion-based models, drawing inspiration from thermodynamic diffusion processes. In particular, we introduced key variants such as DDPM, cDDPM, and DDIM, and reviewed several representative works that apply these models to PET image reconstruction. While diffusion models offer advantages in terms of stable training and high-fidelity generation, they also come with challenges such as high computational cost, slow inference times, and

difficulties in handling high-resolution and 3D images. Overall, the literature reveals a diverse array of strategies—from traditional mathematical and statistical approaches to deep learning frameworks and innovative diffusion-based techniques—all aimed at enhancing the quality of PET image reconstruction. Each method brings its own strengths and limitations, and together they provide valuable insights that inform the development of our proposed approaches.

In this chapter, we have comprehensively outlined the significance of PET image reconstruction, emphasizing the challenges imposed by the inherent properties of PET images. We critically reviewed traditional and deep learning-based methods, highlighting their respective shortcomings in handling noise, preserving fine details, and maintaining global consistency. Most importantly, we examined the current research foundation in diffusion-based methods and identified several key challenges, including slow inference speeds, suboptimal performance, lack of spatial continuity, struggle to deal with multi-dose LPET, and the need for additional modality data. These findings set the stage for the remainder of this thesis, where we will first propose the WiD-PET method to address some of these limitations and subsequently introduce an extended framework CWD-PET that further tackles the multi-dose LPET challenge and enhances reconstruction performance.

Specifically, in Chapter 3, we will provide a detailed problem statement and describe the evaluation metrics and experimental protocols used in PET image reconstruction.

In Chapter 4, we will present our first proposed method, WiD-PET, along with its theoretical underpinnings and experimental validation. The WiD-PET is designed to tackle the slow inference time, lacking-of spatial consistency, and suboptimal performance.

Finally, in Chapter 5, we will introduce our extended approach CWD-PET designed to tackle additional challenges, thereby establishing a comprehensive framework for advanced PET image reconstruction. Therefore, we can further explore multi-dose reconstruction in a single model with the information from cross-modal input.

Chapter 3

Background Introduction of Low Dose PET Reconstruction using Diffusion Techniques

In this chapter, we will introduce the background of the LPET reconstruction task using diffusion techniques, including the problem statement, the current widely accepted evaluation methodologies, and the dataset information.

3.1 Problem Statement

In this section, we briefly outline the problem statement for LPET reconstruction by defining the task of transforming LPET images to SPET images, and we also discuss advanced reconstruction scenarios that involve multi-modal and multi-dose factors.

3.1.1 Reconstruction Task from LPET to SPET

In a typical PET reconstruction task—often also referred to as a denoising or PET reconstruction task—the goal is to transform low-dose PET (LPET) images into standard-dose PET (SPET) images using a learning-based framework. In such tasks, paired datasets consisting of LPET images and their corresponding SPET ground truth images are provided. The framework is designed to learn the mapping between the LPET and SPET images so that it can effectively reconstruct a high-quality SPET

image from a given LPET input. The closer the reconstructed image is to the ground truth SPET image, the better the performance of the reconstruction model.

Formally, let the target SPET image set be defined as $\mathcal{X}^G = \{\mathbf{x}_i^G\}_{i=1}^N$, and the corresponding LPET image set as $\mathcal{X}^L = \{\mathbf{x}_i^L\}_{i=1}^N$.

Specifically, $\mathbf{x}_i^G \in \mathbb{R}^{H\times W\times C}$ represents a SPET image with dimensions $H \times W \times C$, while $\mathbf{x}_i^L \in \mathbb{R}^{H\times W\times C}$ denotes an LPET image of the same size. For diffusion-based methods that address 2D PET reconstruction, each index i corresponds to a PET slice; the SPET image $\mathbf{x}_i^G$ is the standard imaging result that directly corresponds to the LPET image $\mathbf{x}_i^L$ for the same slice.

In diffusion model-based PET reconstruction, the process typically involves two stages. In the first stage, the target SPET image undergoes a forward diffusion process in which noise is gradually added over multiple steps, ultimately transforming the image into a nearly pure noise distribution. In the subsequent stage, a reverse denoising process is employed, where the LPET image is used as a conditioning signal. This conditioning helps to guide the reverse process, fusing the structural and contextual information from the LPET image with the noisy image so that the model can gradually reconstruct an image that closely resembles the original SPET image.

Let us denote the diffusion model-based reconstruction process operator as $D(\cdot \mid \cdot)$. Then, for a given PET slice, the reconstruction process can be formulated as

$$\hat{s}_i = D(\mathbf{x}_i^G \mid \mathbf{x}_i^L). \tag{3.1}$$

where $\hat{s}_i$ represents the reconstructed (or predicted) SPET image for the ith slice, $\mathbf{x}_i^L$ serves as the conditioning LPET image, $\mathbf{x}_i^G$ is the SPET ground truth, and i denotes the index of the PET slice. The set of all reconstructed images is denoted as $\hat{S} = \{\hat{s}_i\}_{i=1}^N$.

This formulation succinctly encapsulates the essence of the diffusion-based PET reconstruction task: the objective is to generate a reconstructed

image $\hat{s}_i$ that approximates the ground truth $\mathbf{x}_i^G$, guided by the corresponding low-dose information provided by $\mathbf{x}_i^L$, and to produce a high-quality set $\hat{S}$ for the entire dataset.

3.1.2 LPET Reconstruction Task with Multi-Modal Data

In addition to the general scenario described in the previous section, there exist more advanced LPET reconstruction tasks that involve multi-modal data inputs. Many diffusion-based PET reconstruction works leverage U-Net architectures as denoising models due to their capability to incorporate and learn from various modalities. For instance, additional modality input—such as prompt guidance information derived from textual cues—can be processed through an encoder or other feature extraction module to yield a representation. Let us denoted the prompt condition by p_i (with the overall set defined as $\mathcal{P} = \{p_i\}_{i=1}^N$).

We then extend the reconstruction operator to accept this multi-modal condition and denote the improved diffusion model as $\hat{D}(\cdot \mid \cdot, \cdot)$. Thus, for a given PET slice, the multi-modal guided reconstruction process can be formulated as

$$\hat{s}_i = \hat{D}(\mathbf{x}_i^G \mid \mathbf{x}_i^L, p_i). \tag{3.2}$$

where $\hat{s}_i$ represents the reconstructed (or predicted) SPET image for the ith slice, $\mathbf{x}_i^L$ serves as the conditioning LPET image, $\mathbf{x}_i^G$ is the SPET ground truth, and p_i encapsulates the encoded additional modality information for the ith slice.

In this formulation, the operator $\hat{D}(\cdot \mid \cdot, \cdot)$ is designed to leverage both the low-dose information from $\mathbf{x}_i^L$ and the supplementary guidance from p_i to produce a more accurate reconstruction of the SPET image or for more complex scenarios.

3.1.3 Multi-dose LPET Reconstruction

In addition to the previously described scenario, another advanced task involves the reconstruction of LPET images acquired at multiple dose

levels. This task requires the implementation and training of a robust model that is not limited to reconstructing images from a single dose level but can efficiently and effectively handle PET reconstructions across various dose levels. Such a capability is of great practical significance in clinical applications, where PET imaging protocols can vary widely. More detailed information on this multi-dose reconstruction task will be provided in Chapter 5.

3.1.4 Evaluation Metrics

To comprehensively assess the performance of our PET reconstruction framework, we employ several quantitative evaluation metrics. The primary metrics used for 3D image quality assessment are:

- **Peak Signal-to-Noise Ratio (PSNR):** PSNR quantifies the overall similarity between the reconstructed image and the ground truth in terms of pixel intensity. A higher PSNR generally indicates better reconstruction quality.
- **Structural Similarity Index Measure (SSIM):** SSIM evaluates the structural and perceptual similarity between images, reflecting how well the reconstruction preserves important structural information.
- **Normalized Mean Squared Error (NMSE):** NMSE provides a normalized measure of the error between the reconstructed and the ground truth images, facilitating a fair comparison across different images and conditions.

In addition to these core metrics, we also consider other evaluation criteria that focus on the preservation of fine image details, such as texture and edge accuracy. These additional metrics will be discussed in greater detail in Chapter 4.

3.2 Datasets Introduction

In this section, we briefly introduce the commonly used dataset in LPET reconstruction—the Ultra Low-Dose PET Challenge (UDPET) dataset [59].

The UDPET dataset comprises a total of 560 ^{18}F-FDG PET scans, along with their corresponding low-dose images acquired at varying time reductions. The dataset was collected from two major sources: 230 subjects scanned with Siemens equipment and 330 subjects scanned with United Imaging equipment. This multi-scanner acquisition ensures that the dataset reflects a broad range of clinical protocols and hardware variability, thereby enhancing the generalizability of reconstruction models.

Each subject's whole-body scan is provided at a resolution of $369 \times 369 \times 644$. For the Siemens data, the pixel spacings are [1.65 mm, 1.65 mm, 1.65 mm], whereas the United Imaging data features pixel spacings of [1.667 mm, 1.667 mm, 2.886 mm]. These differences in spatial resolution and pixel spacing are important factors that are considered during the reconstruction process, particularly when evaluating performance across different scanner types and clinical protocols.

For our work—focused on the clinical value of brain PET imaging—a brain dataset was extracted from the whole-body scans. The brain region was segmented using established methods, yielding brain images with a resolution of $128 \times 128 \times 128$ while retaining the same pixel spacing as the original data. This brain subset is of particular importance, as it enables a targeted evaluation of reconstruction performance in regions where fine anatomical details are critical for clinical diagnosis.

Overall, while the UDPET dataset encompasses whole-body scans, our study primarily leverages the brain dataset to assess and validate PET reconstruction methods. This focused approach not only aligns with the high clinical relevance of brain PET imaging but also provides a robust foundation for our experiments. Detailed preprocessing, segmentation procedures, and dataset splits will be further discussed in subsequent chapters.

Admittedly, other PET datasets exist in the literature. However, within the scope of this study, we employ the UDPET dataset due to its status as the most widely used public benchmark for standard-dose PET reconstruction. Many recent works in the field have also focused on this

dataset alone, enabling consistent comparison and benchmarking. Future work will consider extending the evaluation to additional datasets to validate the generalizability of the proposed methods further.

Chapter 4

Standard-Dose PET Reconstruction from Low-Dose PET by Wavelet-Informed Diffusion Model with Fast Inference

In this chapter, we propose a novel wavelet-informed diffusion model that significantly enhances the reconstruction quality of SPET images while preserving fine details and textures. Moreover, our method reduces the inference time to only 10% of that required by the original DDPM framework. We validate the proposed approach through extensive experiments on LPET images at three different dose levels, including ultra-low-dose data, from the UDPET challenge dataset. The experimental results clearly demonstrate the effectiveness and efficiency of our method.

4.1 Motivations and Contributions

As we mentioned in chapter 2 and chapter 3. Positron emission tomography (PET) plays a critical role in assessing human metabolism and guiding clinical decisions. However, high-quality PET imaging typically requires the injection of substantial radiotracer doses, which increases the

risk of radiation exposure. Consequently, low-dose PET (LPET) reconstruction has become an important research area. Yet, LPET images suffer from a reduced signal-to-noise ratio, leading to potential diagnostic inaccuracies—especially at ultra-low dose levels (e.g., 1/50 or 1/100 of standard dose).

Several methods have been developed to address these challenges, including traditional methods, deep learning methods, and, more recently, diffusion-based methods. With continuous advances in research, traditional approaches have gradually been outperformed by deep learning and diffusion-based reconstruction techniques. Today, high-quality LPET reconstruction tasks are mainly divided into two categories: GAN-based methods and diffusion-based methods.

GAN-based approaches, for example, leverage adversarial training and specialized generator-discriminator architectures to generate high-fidelity images. Studies such as those by Zhao et al. [72] and Zhou et al. [74] have successfully applied cycle-GAN frameworks for LPET reconstruction. However, the inherent instability of adversarial training often leads to issues such as mode collapse, resulting in limited discriminability and reduced reliability in clinical contexts.

More recently, diffusion-based models like DDPM [54] have shown promise by leveraging a forward noise-injection process followed by a reverse denoising process. Despite their advantages in stable training and high-fidelity image generation, these models face several significant challenges when applied to PET reconstruction. **First**, their iterative sampling procedure is computationally expensive and results in slow inference speeds, which is particularly problematic for high-resolution images. **Second**, the conventional diffusion process tends to uniformly denoise in the image space, often failing to recover fine details—especially in high-frequency components such as textures and edges in ultra-low-dose images. **Third**, because most diffusion models are originally designed for 2D image processing, they handle PET slices independently, leading to a lack of spatial continuity across 3D reconstructions.

These challenges motivate the need for an improved approach. To

address the limitations of existing diffusion-based models, we propose a novel framework—termed the Wavelet-informed Diffusion Model with Fast Inference (WiD-PET). The key motivations and novelties we contributed behind WiD-PET are threefold:

1. **Enhanced Detail Recovery:** By incorporating wavelet transformations and a high-frequency detail enhancer module, our model is designed to better recover fine image details and textures, thus improving the quality of reconstruction.
2. **Accelerated Inference:** Processing in the wavelet domain not only emphasizes critical image features but also reduces the scale of the input image to the framework, thereby cutting inference time to approximately 10% of that required by the original DDPM framework.
3. **Spatial Consistency:** To overcome the challenge of slice-wise processing in 2D diffusion models, WiD-PET integrates a Spatial Consistency Feature Extractor (SCFE) and Spatial Consistency Attention (SCA) mechanism, ensuring seamless continuity across 3D PET slices, and thus improve the quality of the reconstructed SPET images.

Extensive experiments conducted on LPET images at various dose levels (1/20, 1/50, and 1/100) validate the effectiveness of WiD-PET, demonstrating superior reconstruction quality, enhanced detail preservation, and markedly improved inference efficiency. In the following chapters, we will further elaborate on the design of WiD-PET and provide a detailed evaluation of its performance in comparison with existing methods.

4.2 Methodology

In this section, we introduce WiD-PET, a wavelet-based diffusion framework with efficient inference designed for PET reconstruction. Our proposed approach is structured into several subsections to facilitate a clear understanding of its components and underlying principles.

- In Section 4.2.1, we present the detailed preliminaries essential for this chapter, including new concepts and specific notations that underpin our methodology.
- Section 4.2.2 provides an overview of the proposed framework, outlining its main components and the overall workflow.
- In Section 4.2.3, we describe the fast wavelet-informed diffusion model in detail. This section covers the design of a 2D-DWT-IDWT module and a spatial-consistency-informed denoising model, which incorporates key building blocks such as the Spatial Consistency Feature Extractor (SCFE) and Spatial Consistency Attention (SCA) modules, as well as a High-frequency Component Enhancer (HFE).
- Finally, Section 4.2.4 explains the combined loss function that has been designed to ensure high-quality reconstruction of LPET images.

This structured approach not only clarifies the individual components of WiD-PET but also illustrates how they integrate to achieve efficient and high-quality PET reconstruction.

4.2.1 Preliminaries

In this section, we briefly review the fundamental concepts behind diffusion models and wavelet transforms, both of which play key roles in our proposed framework. Many of the symbols and definitions introduced here have been previously presented in the literature review; for brevity, we restate only the essential components.

Denoising Diffusion Probabilistic Models (DDPM) can provide a robust framework for generating high-quality reconstructions by learning to reverse a gradual noise-corruption process [32]. In a standard DDPM, two stages are involved: the forward diffusion process and the reverse denoising process.

For our PET reconstruction task, let the target SPET image set be defined as $\mathcal{X}^G = \{\mathbf{x}_i^G\}_{i=1}^N$, and the corresponding LPET image set as $\mathcal{X}^L = \{\mathbf{x}_i^L\}_{i=1}^N$.

Here, each $\mathbf{x}_i^G \in \mathbb{R}^{H\times W\times C}$ represents a SPET image, and $\mathbf{x}_i^L \in \mathbb{R}^{H\times W\times C}$ denotes the corresponding LPET image for the ith PET slice. In our framework, the generation of $\mathbf{x}_i^G$ is conditioned not only on $\mathbf{x}_i^L$ but also on additional spatial sequence information, collectively represented as $\mathcal{C} = \{\mathbf{c}_i\}_{i=1}^N$. We also denote the timestep by t, with $\mathbf{x}_t^G$ representing the noisy SPET image at timestep t and $\mathbf{x}_0^G$ the noise-free image.

Forward Diffusion Process: The forward process gradually adds noise to the SPET images, which can be modeled as

$$q(\mathbf{x}_t^G \mid \mathbf{x}_{t-1}^G) = \mathcal{N}(\mathbf{x}_t^G; \sqrt{\alpha_t}\,\mathbf{x}_{t-1}^G, \beta_t\mathbf{I}). \tag{4.1}$$

where β_t denotes the noise increment at timestep t and the scaling factor is defined as $\alpha_t = 1 - \beta_t$.

Reverse Denoising Process: The reverse process iteratively reconstructs the clean image by removing noise:

$$p_\theta(\mathbf{x}_{t-1}^G \mid \mathbf{x}_t^G, \mathbf{c}) = \mathcal{N}(\mathbf{x}_{t-1}^G; \mu_\theta(\mathbf{x}_t^G, t, \mathbf{c}), \Sigma_\theta(\mathbf{x}_t^G, t, \mathbf{c})). \tag{4.2}$$

where μ_θ and Σ_θ are the predicted mean and variance, respectively, conditioned on the auxiliary information $\mathbf{c}$.

Wavelet Transformation is a powerful tool for decomposing an input image into multiple frequency bands [70]. In our framework, we define the 2D wavelet transformation as $f_{\text{dwt}}(\cdot)$. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$, the transformation decomposes the image into a set of four components:

$$f_{\text{dwt}}(\mathbf{x}) = \{\mathbf{LL}, \mathbf{LH}, \mathbf{HL}, \mathbf{HH}\}. \tag{4.3}$$

where $\mathbf{LL}, \mathbf{LH}, \mathbf{HL}, \mathbf{HH}$ denote the decomposed components. Specifically, the component $\mathbf{LL} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ represents the **low-frequency** result, capturing the global structure of the image. In contrast, the components $\mathbf{LH}, \mathbf{HL}, \mathbf{HH} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ collectively form the **high-frequency** results, which capture horizontal, vertical, and diagonal details, respectively.

By decomposing the image using the 2D wavelet transform, the spatial dimensions are significantly reduced, thereby decreasing computational cost and reducing computation time. Moreover, this transformation effectively separates the primary structure (low frequency) from the fine details (high frequency), which lays a solid foundation for subsequent detail reconstruction.

4.2.2 Overview

An overview of the proposed WiD-PET framework is illustrated in Figure 4.1. First, the SPET images, along with the LPET condition and its adjacent slices (collectively referred to as the spatial sequence), are decomposed using the Haar 2D wavelet transform (2D-DWT). This decomposition yields the low-frequency component $\mathbf{LL}_x$ and the high-frequency components $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_x$, where the subscript $x \in \{g, c, S\}$ indicates the ground-truth SPET image (g), the LPET condition (c), and the spatial sequence from LPET (S), respectively.

Next, the low-frequency component $\mathbf{LL}_g$ is processed through a forward diffusion process. The resultant noisy representation is then combined with the corresponding low-frequency condition $\mathbf{LL}_c$ and fed into the denoising model. Concurrently, the low-frequency component $\mathbf{LL}_S$ extracted from the LPET-derived spatial sequence is also used as a condition to provide spatial consistency guidance. This guidance is refined through a Spatial-Consistency Feature Extractor (SCFE) block followed by a dedicated Spatial Consistency Attention (SCA) module, which effectively steers the denoising process.

Meanwhile, the high-frequency components are further enhanced using a High-frequency Enhancer to restore fine details and textures that are crucial for accurate reconstruction. Finally, the enhanced high-frequency components and the reconstructed low-frequency components are fused via the inverse 2D wavelet transform (2D-IDWT) to produce the final reconstructed images.

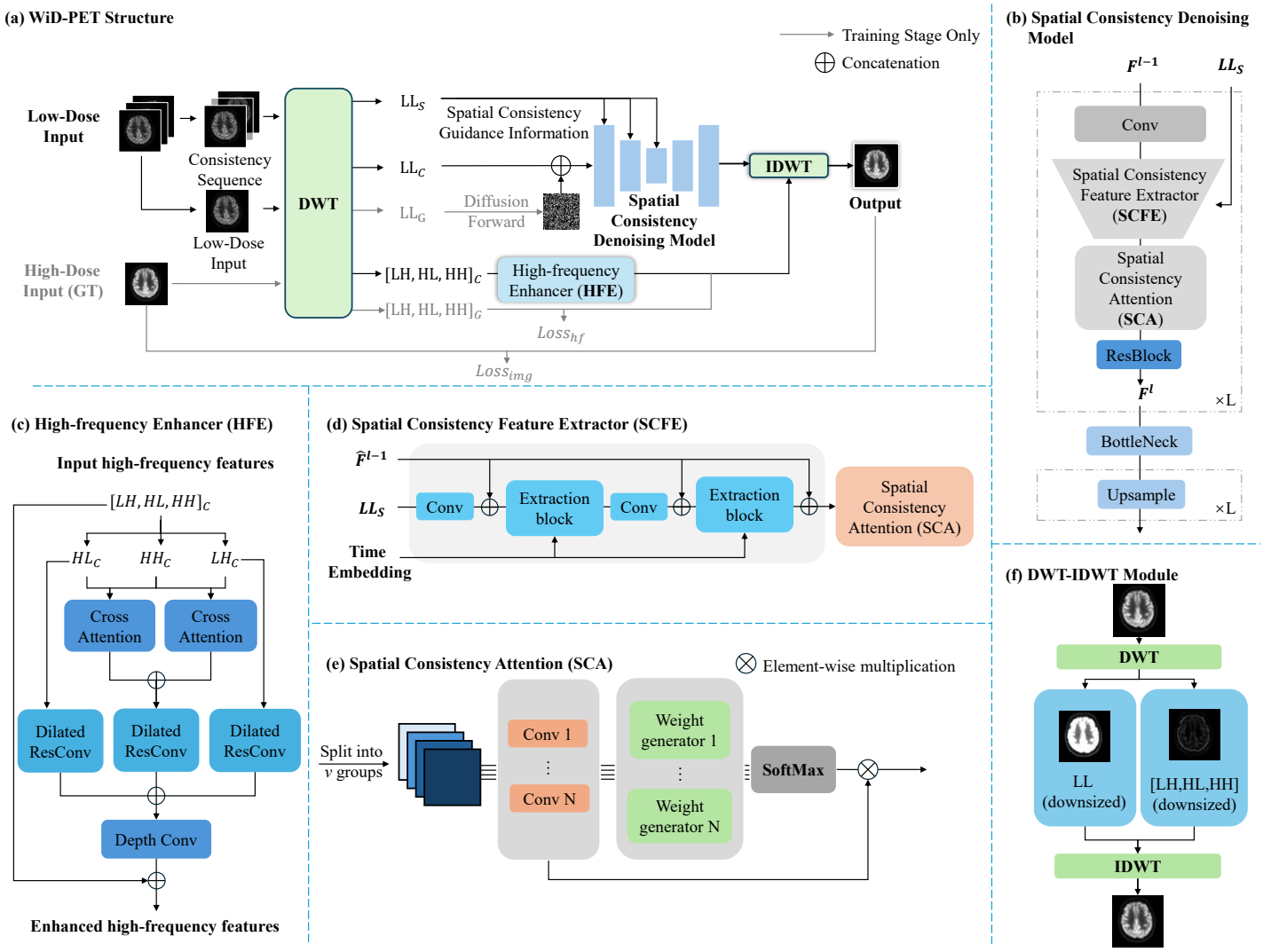


Figure 4.1: Overview of WiD-PET (a): The SPET, LPET, and spatial sequence (adjacent slices) are decomposed into low- and high-frequency components using the wavelet transform in the DWT-IDWT module (b). The low-frequency components are processed by a diffusion-based denoiser (c), while the high-frequency components are enhanced using the high-frequency enhancer (Fig. 4.3) to restore global contrast and fine details. The outputs are then recombined and inversely transformed to produce the final reconstructed SPET image.

4.2.3 Fast Wavelet-informed Diffusion Architecture

Based on the existing challenges—particularly the low computational efficiency in reconstructing full-dose SPET images from LPET images and the unsatisfactory restoration of fine image details—we propose a fast wavelet-informed diffusion architecture that leverages 2D wavelet transformation. This novel approach is designed to accelerate the reconstruction process while enhancing the recovery of critical image features.

DWT-IDWT Module

As aforementioned, our approach employs a DWT-IDWT module (illustrated in Fig. 4.1(b)) to decompose the input SPET ground truth, LPET condition, and the corresponding spatial sequence S into their respective frequency components. Specifically, these inputs are decomposed into the low-frequency components $\mathbf{LL}_g$, $\mathbf{LL}_c$, and $\mathbf{LL}_S$ and the high-frequency components $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_g$, $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_c$, and $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_S$. After decomposition, the spatial dimensions of each component are reduced from $C \times H \times W$ to $C \times \frac{H}{2} \times \frac{W}{2}$, effectively reducing the overall input size by a factor of 4.

The low-frequency component of the SPET image, $\mathbf{LL}_g$, undergoes a forward diffusion process and is then merged with the low-frequency component of the LPET image, $\mathbf{LL}_c$. This combined representation is subsequently fed into the denoising model to reconstruct the global contrast of the image. In parallel, the high-frequency components of the LPET input are processed by the High-frequency Component Enhancer to restore fine details and textures. Finally, the outputs from the denoising and enhancement stages are fused using the inverse 2D wavelet transform (IDWT) to generate the final reconstructed image.

Spatial-consistency Informed Denoising Model

The denoising model learns to reconstruct noise-free images from the inputs of the low-frequency components. It adopts a self-attention U-Net architecture with spatial consistency guidance in encoding input

features, as shown in Figure 4.1 (c). As seen, the model consists of L encoder layers, the bottleneck layer, and L decoder layers (upsampling). Compared with the conventional U-Net architecture used in DDPM, our denoising model introduces two innovative modules into each encoder layer, i.e., the Spatial Consistency Feature Extractor (SCFE) and the Spatial Consistency Attention (SCA). They are used to effectively leverage the adjacent slices to address the limitations of existing 2D-based diffusion and GAN models in capturing spatial continuity and consistency during training and inference. Specifically, each encoder layer consists of a convolution block, the SCFE module, the SCA module, and the Resblock. The bottleneck layer and the decoder layer are the same as those used in DDPM. The SCFE and SCA modules are elaborated as follows.

Spatial Consistency Feature Extractor (SCFE) takes three inputs: first, the output from the last encoder layer $\hat{F}^{l-1} = \text{CONV}(F^{l-1})$, second, the spatial consistency guidance $\mathbf{LL}_S$, and third, the time embedding of the current time step. At the beginning, when $l = 1$, $F^{l-1} = F^0 = \text{CONCAT}(\mathbf{LL}_\mathbf{g}^{\mathbf{noise}}, \mathbf{LL}_\mathbf{c})$. Here $\mathbf{LL}_\mathbf{g}^{\mathbf{noise}}$ is the noise-corrupted version of LL_g. The architecture of SCFE, illustrated in Fig. 4.2 (a), employs a residual convolutional network for processing. To begin, the spatial sequence LL_S is aligned with F^{l-1} using a 1×1 convolution and then concatenated with $\hat{F}^{l-1}$). A specialized extraction block, consisting of multiple convolutional layers, extracts spatial consistency information by performing feature fusion and iterative noise adjustment. This extraction operation is applied twice within the SCFE to ensure robust feature refinement. At the final stage, the extracted spatial consistency features are integrated with the original input feature $\hat{F}^{l-1}$ through a residual connection, which helps preserve the input's integrity while enhancing it with the extracted spatial information. The resulting features are then passed to the corresponding attention module for further processing.

Spatial Consistency Attention (SCA) enhances spatial feature extraction by leveraging multi-scale convolutional kernels and normalized attention weights. The structure of SCA is illustrated in Fig. 4.2 (b). The input spatial features are first divided into v subgroups along the channel dimension, where v corresponds to the number of slices in the spatial

sequence. Each subgroup is then processed by convolutional blocks with kernels of varying sizes to capture multi-resolution features effectively. The outputs of these convolutions are passed through a global pooling layer, followed by a weight generator consisting of fully connected layers, and finally normalized using a SoftMax layer to produce attention weights. The normalized attention weights are applied to the corresponding sub-groups of the convolutional outputs, and the weighted features are subsequently multiplied and refined. The weighted outputs from all sub-groups are concatenated to form the final output of the module.

High-frequency Enhancer

The high-frequency enhancer (HFE) is a residual convolutional structure designed to improve the recovery of fine details in LPET images. It processes the high-frequency components of LPET, denoted as $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_c$. At the beginning of HFE, the paired components $[\mathbf{LH}, \mathbf{HL}]_c$ and $[\mathbf{HL}, \mathbf{HH}]_c$ are processed through cross-attention blocks to capture the interactions between these high-frequency features (capturing complementary directional information). Next, the individual components $\mathbf{LH}_c$ and $\mathbf{HL}_c$, together with the concatenated outputs from the cross-attention blocks, are fed into dilated convolution blocks to extract multi-scale contextual information. The resulting feature maps are then concatenated and passed through a depthwise convolution layer. Finally, these features are fused with the original high-frequency inputs $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_c$ via a residual connection. The overall structure of the HFE is illustrated in Fig. 4.3. This design effectively integrates information across different high-frequency components, thereby enhancing the model's ability to reconstruct fine details in PET images from LPET inputs.

4.2.4 Loss Function

In order to better tackle the task, We design a comprehensive loss function to optimize our proposed method by incorporating multiple components that address various aspects of the reconstruction process. The noise loss, $\mathcal{L}_{\text{noise}}$, minimizes the discrepancy between the Gaussian noise

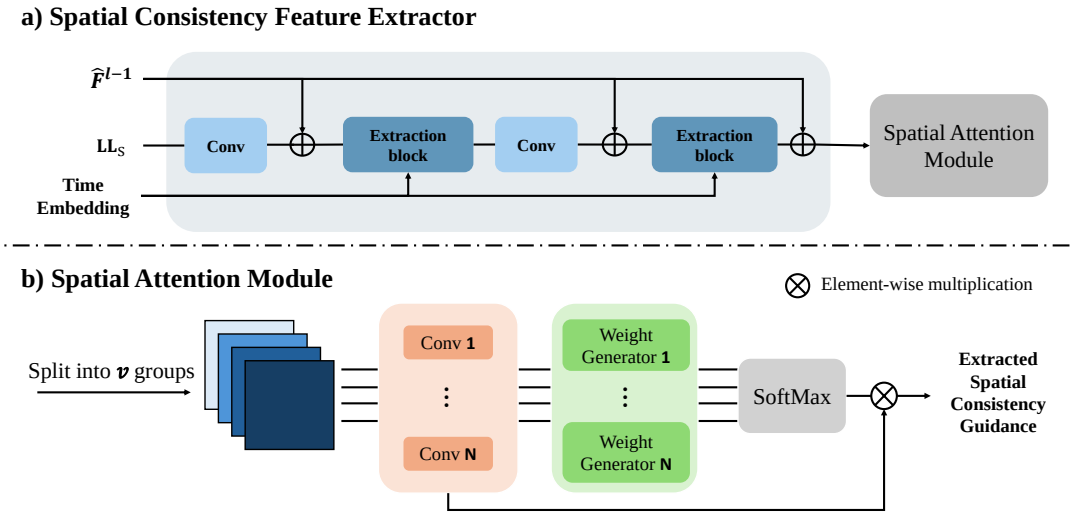


Figure 4.2: (a) **Spatial Consistency Feature Extractor (SCFE):** This module takes the output from the previous layers $\hat{F}^{l-1}$, the low frequency components of the spatial sequence $\mathbf{LL}_S$, and the current time embedding as the inputs. The extracted spatial features are sent to the **SCA** module. (b) **Spatial Consistency Attention (SCA) Module:** The input features are divided into v sub-groups along the channels, where v is equal to the number of slices in the spatial sequence. Each sub-group is processed through convolutions of multi-scale kernels; the extracted multi-scale features are weighted by normalized attention weights within each sub-group; the weighted outputs of subgroups are concatenated as the output of the module.

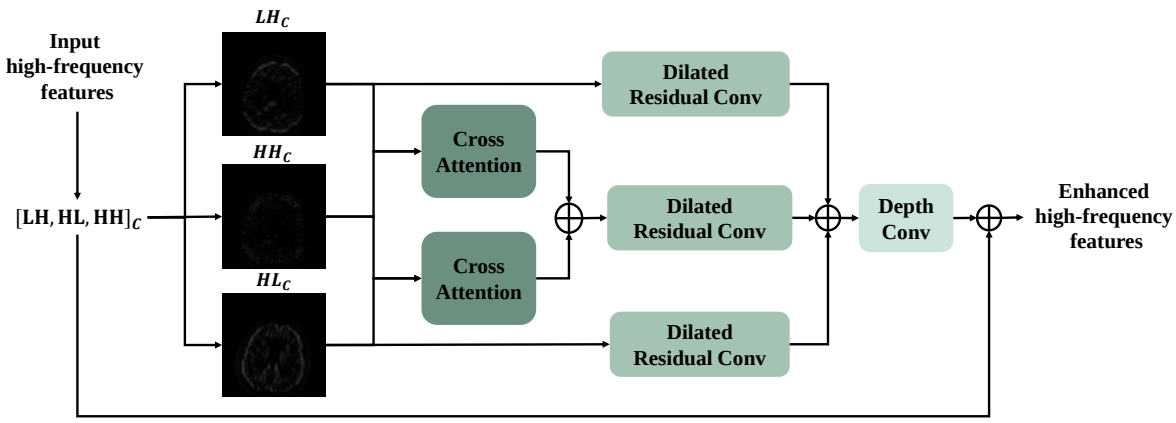


Figure 4.3: The structure of the **High-frequency Enhancer**. This module takes the decomposed high-frequency components of LPET $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_c$ from the DWT module as the inputs, using cross-attention, dilated residual convolution, and depth convolution to enhance high-frequency features.

$\mathbf{n}$ added during the forward diffusion process and the predicted noise $\hat{\mathbf{n}}$ output from the denoising model, and is defined as:

$$\mathcal{L}_{\text{noise}} = \|\mathbf{n} - \hat{\mathbf{n}}\|_2^2. \tag{4.4}$$

To enhance fine details and preserve high-frequency information, we introduce a high-frequency loss, $\mathcal{L}_{\text{hf}}$, which combines an ℓ_2 loss for the high-frequency components with total variation (TV) regularization. This loss is defined as:

$$\mathcal{L}_{\text{hf}} = \lambda_{\ell_2} \|\mathbf{x}_{\text{hf}} - \hat{\mathbf{x}}_{\text{hf}}\|_2^2 + \lambda_{\text{tv}} \, \text{TV}(\hat{\mathbf{x}}_{\text{hf}}). \tag{4.5}$$

where λ_{ℓ_2} and λ_{tv} are weighting factors balancing the ℓ_2 loss and TV regularization, respectively. Here, $\mathbf{x}_{\text{hf}}$ denotes the concatenation of the high-frequency components of the SPET image (serving as the ground truth), and $\hat{\mathbf{x}}_{\text{hf}}$ represents the corresponding high-frequency output produced by our model.

Additionally, we define a reconstruction loss, $\mathcal{L}_{\text{img}}$, using an ℓ_1 loss to measure the difference between the original SPET image $\mathbf{x}^G$ and the

reconstructed image $\hat{\mathbf{x}}^G$ obtained after applying the 2D inverse discrete wavelet transform (IDWT):

$$\mathcal{L}_{\text{img}} = \|\mathbf{x}^G - \hat{\mathbf{x}}^G\|_1. \tag{4.6}$$

Combining these components, the final loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{hf}} + \mathcal{L}_{\text{img}}. \tag{4.7}$$

4.3 Experiments and Results

In this section, we first describe the implementation details of our WiD-PET model, followed by a comprehensive evaluation of the Ultra-low Dose PET (UDPET) dataset [59]. We analyze the effectiveness of our approach by comparing its performance against several state-of-the-art methods using both qualitative assessments and quantitative metrics. Additionally, we present ablation studies to demonstrate the contribution of each module within our framework.

4.3.1 Implementation Details

Our proposed method leverages ultra-low-dose LPET images acquired at 1/20, 1/50, and 1/100 dose levels, with corresponding normal-dose SPET images (1/1 dose) serving as the ground truth. Both LPET and SPET images are resized to 128×128 for input. The framework is implemented in PyTorch and run on an NVIDIA RTX 3090 GPU. During training, data augmentation techniques such as random cropping and flipping are applied to the input images. The initial learning rate is set to 1×10^{-4} and decays by a factor of 0.8 every 5000 iterations. Furthermore, the 2D-DWT-IDWT module operates with a transformation scale of 2.

4.3.2 Evaluation Metrics

To evaluate the reconstruction quality of our method, we employ PSNR, SSIM, and NMSE for quantitative comparison with baseline models. The

details of PSNR, SSIM, and NMSE are described in previous chapters, here we note that higher PSNR and SSIM values, along with lower NMSE, indicate better performance. In addition, we assess image detail recovery using Gradient Loss—which emphasizes local structures and edges through first-order intensity changes—and Brenner Gradient Loss, which captures intricate textures and sharpness via second-order variations. Finally, the inference time is measured as the average duration required to generate a full 3D SPET image from whole-brain LPET slices.

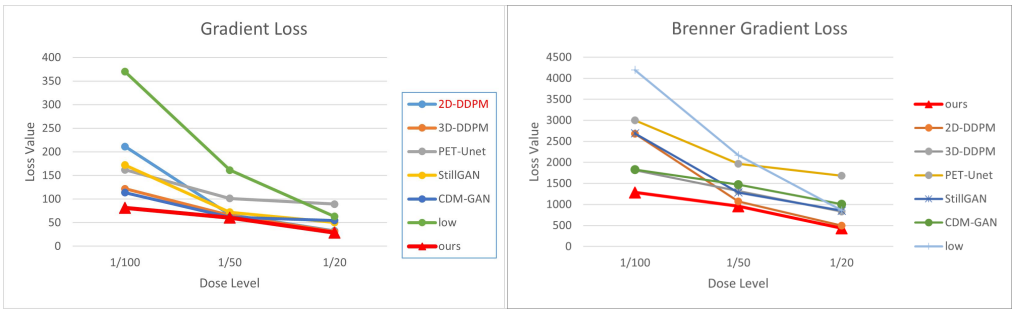
4.3.3 Experimental Results

The quantitative comparison results of our method and the baselines are presented in Table 4.1. We compare our approach with the general diffusion-based model 2D-DDPM [32], and 3D-DDPM [18], the GAN-based model Still-GAN [47], as well as state-of-the-art PET reconstruction methods including CDM-GAN [29] and PET-Unet [11]. These methods were selected due to the availability of publicly released code, ensuring reproducibility. For a fair comparison, we reran the provided implementations on the same training-test partition and demonstrated performance across three different LPET dosage levels. Additionally, we include evaluation metrics for LPET images at varying doses as reference values.

Table 4.1: Comparison of reconstruction at 1/20, 1/50, and 1/100 dose levels on the UDPET dataset.

Methods	1/100 dose			1/50 dose			1/20 dose			Inference Time
	PSNR	SSIM	NMSE ($\times 10^{-4}$)	PSNR	SSIM	NMSE ($\times 10^{-4}$)	PSNR	SSIM	NMSE ($\times 10^{-4}$)	(s/128 slices)
Low dose[59]	15.46	0.46	22.02	21.70	0.70	22.73	22.72	0.77	4.02	–
2D-DDPM[32]	22.68	0.76	7.27	25.86	0.86	5.62	26.11	0.92	3.27	268.00
3D-DDPM(cWDM)[18]	25.16	0.85	14.10	27.40	0.90	4.20	28.81	0.93	6.10	108.05
Still-GAN[47]	23.61	0.83	9.69	24.48	0.85	7.04	25.63	0.90	3.96	26.12
CDM-GAN[29]	23.95	0.84	12.62	26.80	0.85	3.80	28.84	0.91	3.12	25.24
Pet-Unet[11]	23.30	0.81	8.43	24.33	0.84	6.29	27.93	0.91	2.60	**2.02**
WiD-PET(Ours)	**26.68**	**0.89**	**3.60**	**27.82**	**0.91**	**3.11**	**29.93**	**0.94**	**1.85**	<u>20.24</u>

Overall Quality As shown in Table 4.1, all comparison methods significantly improve the quality of LPET images. Among them, our proposed method consistently achieves superior performance across 1/20, 1/50,



(a) Gradient Loss Comparison (b) Brenner Gradient Loss Comparison

Figure 4.4: Comparison of recovery of details by (a) Gradient Loss and (b) Brenner Gradient Loss across LPET, 2D-DDPM, 3D-DDPM, PET-Unet, STillGAN, CDM-GAN, and our proposed method at varying dose levels.

and 1/100 dose levels. Taking the 1/100 dose level as an example, our WiD-PET significantly enhances the PSNR/SSIM/NMSE values from 15.46/0.46/22.02 of LPET to 26.68/0.89/3.60 in the reconstructed SPET images. This represents a substantial improvement over the second-best performer, 3D-DDPM, which achieves 25.16/0.85/14.10. Meanwhile, although the general-purpose diffusion-based model DDPM outperforms the GAN-based StillGAN, both methods fall short compared to the specialized PET reconstruction models CDM-GAN and Pet-Unet. In contrast, our WiD-PET, tailored for PET reconstruction, demonstrates superior performance at enhancing high-frequency details while leveraging spatial consistency across slices to reduce artifacts and improve visual smoothness.

Detail Recovery To assess high-frequency detail preservation, we evaluate **gradient loss** and **Brenner gradient loss**, which measure local structure and edge recovery via first- and second-order differences. As shown in Fig. 4.4, WiD-PET achieves the lowest losses across all dose levels. At moderate doses (1/20 and 1/50), diffusion-based models (2D-DDPM, 3D-DDPM, and WiD-PET) outperform GAN-based methods in capturing fine details. However, at the ultra-low dose (1/100), both vanilla 2D- and 3D-DDPM struggle—especially 2D-DDPM, which shows a significantly higher loss. In contrast, WiD-PET's tailored strategies substantially enhance detail recovery.

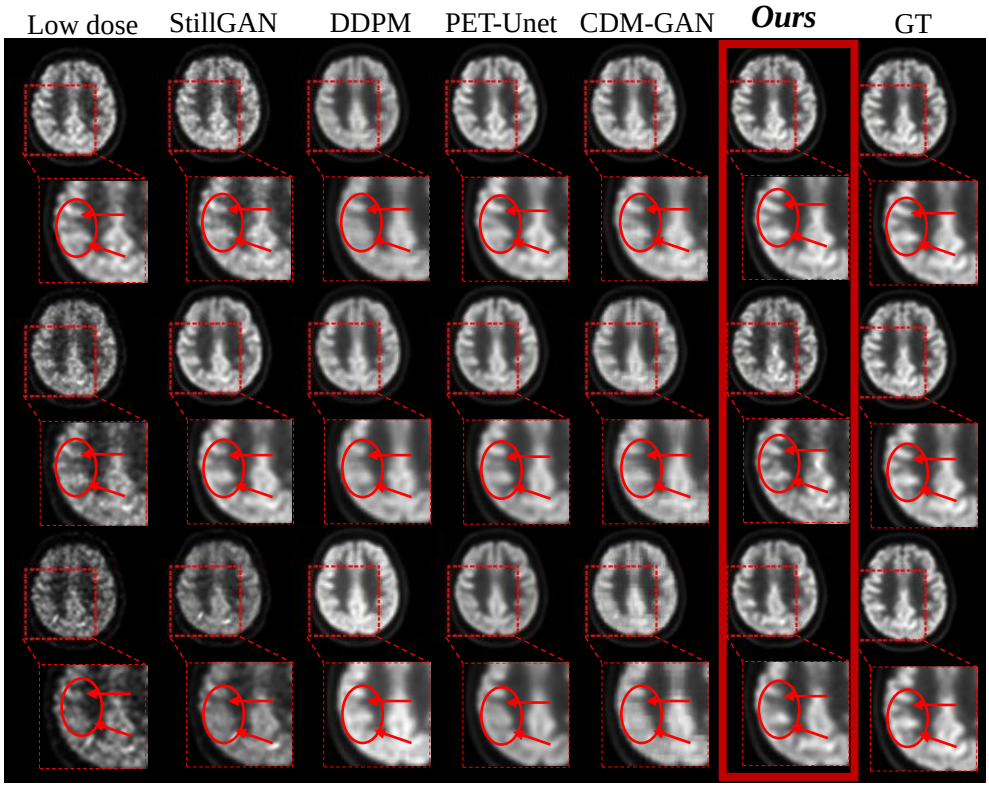


Figure 4.5: Visual comparison of 2D methods' reconstruction results across different dose levels. For each dose level, the first row presents the full-slice view, while the second row provides a zoomed-in view of the regions highlighted by red boxes in the first row.

Visual Comparison A visualization comparison of the reconstruction results is given in Figure 5.5. As can be seen, the images reconstructed by our WiD-PET model are sharper, carrying richer details that resemble the ground truth. Our advantages become more salient at 1/100 dose level. This is consistent with the quantitative analysis, verifying the effectiveness of our proposed strategies.

Inference Speed Beyond its superior reconstruction performance, our WiD-PET demonstrates remarkable efficiency during inference. As shown in Table 4.1, the integration of wavelet transform reduces the inference time by nearly 10-fold, from 268s per 128 slices for 2D-DDPM to just 20.24s per 128 slices (5 times less than 3D-DDPM model with wavelet decomposition module). This speed surpasses both GAN-based models

and is second only to Pet-Unet, which, despite its faster inference, delivers significantly inferior reconstruction quality.

4.3.4 Ablation Study

We conducted ablation studies to evaluate the effectiveness of each component in our method, with the results summarized in Table 4.2. Here, the "Base Model" refers to the vanilla 2D-DDPM model. As shown in Row 2, when we employ the wavelet transform to decompose the input images into low- and high-frequency components and enhance the high-frequency components using our proposed High-frequency Component Enhancer, the PSNR/SSIM values improve significantly, the quantitative results increased from 22.68/0.76, 25.86/0.86, and 26.11/0.82 to 25.19/0.79, 25.26/0.86, and 28.33/0.71 at dose levels 1/100, 1/50, and 1/20, respectively. These results underscore the importance of treating high-frequency components specially in diffusion-based models for UDPET reconstruction.

In Row 3, we introduce spatial-consistency guidance by incorporating both the SCFE and SCA modules, which must be used in tandem. This further enhances performance to 25.41/0.79, 25.28/0.79, and 28.58/0.84, demonstrating the benefit of capturing spatial consistency during the denoising process. In the subsequent row, the addition of our specially designed high-frequency loss term $\mathcal{L}_{\text{hf}}$ further boosts performance to 25.91/0.86, 26.16/0.87, and 28.69/0.92, validating its superior ability to correct high-frequency details compared to conventional loss terms in standard diffusion models.

By integrating all of these components, our full method achieves its best reconstruction performance across all dose levels, thereby confirming the contribution of each module to the overall success of our approach.

Table 4.2: Results of ablation experiments on both dose levels from the UDPET dataset. We removed the wavelet transformation components, SCFE, and high-frequency loss from the base diffusion model to assess the performance impact.

Components				1/100 dose			1/50 dose			1/20 dose		
Base Model	Wavelet Transform and Enhancer	Spatial Guidance	High-frequency Loss	PSNR	SSIM	NMSE ($\times 10^{-4}$)	PSNR	SSIM	NMSE ($\times 10^{-4}$)	PSNR	SSIM	NMSE ($\times 10^{-4}$)
✓	·	·	·	22.68	0.76	7.27	25.86	0.86	5.62	26.11	0.82	3.27
✓	✓	·	·	25.19	0.79	6.50	25.26	0.86	5.47	28.33	0.71	2.30
✓	✓	✓	·	25.41	0.79	5.81	25.28	0.79	5.18	28.58	0.84	2.24
✓	✓	·	✓	25.91	0.86	4.63	26.16	0.87	4.57	28.69	0.92	2.22
✓	✓	✓	✓	**26.68**	**0.89**	**3.60**	**27.82**	**0.91**	**3.11**	**29.93**	**0.94**	**1.85**

4.3.5 Summary

In this chapter, we presented our novel wavelet-informed diffusion PET reconstruction method, WiD-PET, which aims to reconstruct standard-dose PET (SPET) images from low-dose PET (LPET) images. Our approach has demonstrated robust reconstruction performance across three different LPET dose levels, including ultra-low-dose scenarios, by achieving superior image quality with enhanced spatial consistency and finer details, all while significantly reducing inference time compared to conventional diffusion-based models.

However, it is important to note that WiD-PET primarily focuses on single-dose reconstruction and does not incorporate additional multimodal or multi-dose information to further guide the reconstruction process. To address these limitations and further improve reconstruction quality under varied dose conditions, we introduce a novel extension in the next chapter. This new approach leverages extra guidance information to enhance the robustness and accuracy of PET reconstructions, thereby paving the way for more comprehensive and clinically applicable solutions.

Chapter 5

Standard-Dose PET Reconstruction from Multi-dose LPET by Cross-dose Wavelet-informed Refine Diffusion Model

In this chapter, we propose the Cross-dose Refine Wavelet-informed Diffusion (CWD-PET) framework—an extension of our previously introduced WiD-PET model—designed to address advanced cross-dose PET reconstruction tasks. CWD-PET integrates multi-modal input information with multi-dose LPET images to better capture additional structural and contextual cues essential for high-quality reconstruction. To fully exploit this new information and tackle the challenges posed by more complex reconstruction scenarios, we have developed novel modules, and a specialized network architecture tailored for CWD-PET. Comprehensive experiments, conducted on a multi-dose LPET dataset derived from the UDPET dataset (encompassing all low-dose levels), demonstrate the effectiveness and robustness of our approach.

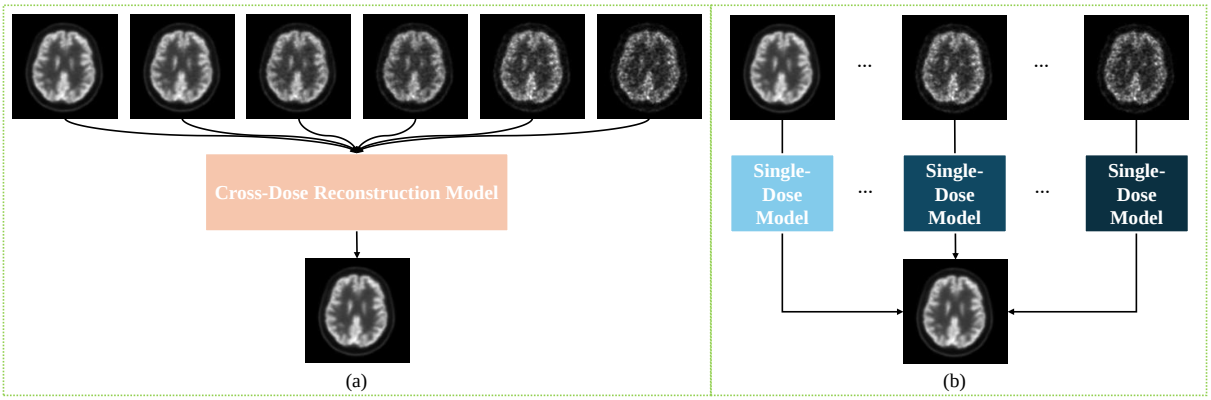


Figure 5.1: An illustration of the differences between single-dose and cross-dose reconstruction models. (a) Cross-dose reconstruction model: a single model capable of handling LPET images acquired at multiple dose levels; (b) Single-dose reconstruction models: distinct models that must be trained separately for each LPET dose level.

5.1 Motivations and Contributions

As discussed in Chapter 4, PET image reconstruction plays a critical role in clinical diagnosis and treatment planning, underscoring the importance of achieving high-quality reconstructions. While numerous studies have proposed GAN-based and diffusion-based methods to optimize reconstruction for a specified LPET dose level [10], these approaches generally overlook a crucial clinical challenge: the lack of a universally accepted optimal LPET dose. In clinical practice, LPET images are acquired at varying dose levels, and methods tailored to a single dose level often fail to generalize across this range. For instance, the gap in image quality between LPET images acquired at 1/2 dose and those at 1/100 dose is significant, with typical PSNR and SSIM values of approximately 28.5 and 0.89 for 1/2 dose, compared to 15.46 and 0.46 for 1/100 dose. This substantial distribution discrepancy hinders the performance of models optimized solely for a single-dose reconstruction task when applied to multi-dose scenarios.

To address this gap, we propose the CWD-PET framework—an extension of our previously introduced WiD-PET model—that integrates multi-modal input information with cross-dose LPET images to tackle

advanced cross-dose PET reconstruction tasks. As illustrated in Figure 5.1, unlike single-dose reconstruction models which require separate training for each LPET dose level, the cross-dose model is designed to handle multiple dose levels with a single, unified framework. Our approach introduces prompt guidance that encapsulates pre- and post-dose level information, which is incorporated as a multi-modal input to guide the reconstruction process. In order to effectively fuse this additional modality data, we have developed a dedicated **Prompt Embedding Fusing module (PEF)**. Furthermore, to enhance the overall reconstruction quality across varying dose levels, we design a lightweight yet effective **Refinement-Net (RFN)** that further improves the output after the inverse DWT operation.

In summary, the key contributions of this work are:

1. **Extending the Wavelet-informed Diffusion framework** to address cross-dose PET reconstruction, an area that has been largely overlooked despite its significant clinical relevance.
2. Introducing a multi-modal input strategy via prompt guidance and a novel **Prompt Embedding Fusing module (PEF)** to integrate pre- and post-dose information into the reconstruction process.
3. Introducing a **lightweight Refinement Network (RFN)** to further enhance the final reconstruction quality after the inverse wavelet transform, ensuring improved detail recovery and smoother image outputs in the complex cross-dose reconstruction task with minimal additional computational cost.

These contributions collectively advance the state-of-the-art in PET image reconstruction and hold promise for significantly improving clinical diagnostic accuracy across diverse LPET dose levels.

5.2 Methodology

In this section, we introduce in detail our newly proposed cross-dose level reconstruction model, CWD-PET. Building upon the work presented in the previous chapter 4, CWD-PET leverages the fast inference and

outstanding reconstruction performance of our wavelet-informed diffusion framework and spatial consistency guidance architectures. In addition, by incorporating novel prompt guidance information, we employ a prompt encoder to encode this supplementary data and introduce a specially designed Prompt Embedding Fusing (PEF) module to seamlessly integrate the new guidance into the denoising model. Furthermore, we design a Refinement-Net (RFN) to further boost reconstruction performance after the inverse discrete wavelet transform (IDWT), thereby enhancing the overall quality of the final output.

The remainder of this section is organized as follows. In Section 5.2.1, we present the necessary preliminaries. Section 5.2.2 provides an overview of the CWD-PET framework. In Section 5.2.3, we detail the design and implementation of the PEF and RFN modules. Finally, Section 5.2.4 describes the combined loss function used to optimize our approach.

5.2.1 Preliminaries

We extend the conventional diffusion-based reconstruction framework by incorporating multi-modal guidance, which enables improved performance in cross-dose PET reconstruction tasks. This section introduces the key components from the CLIP model (the CLIP tokenizer and encoder) that are integral to our method and reviews the fundamental concepts of multi-modal diffusion reconstruction.

CLIP for Multimodal Guidance: To effectively incorporate supplementary multimodal information into our framework, we utilize components from the CLIP model [49]. Let τ_i denote the raw textual prompt for the ith sample, and define the set of all raw prompts as $\mathcal{T} = \{\tau_i\}_{i=1}^{N}$. The CLIP tokenizer is then applied to convert each raw prompt τ_i into a sequence of tokens:

$$\mathbf{tok}_i = f_{\text{token}}(\tau_i). \tag{5.1}$$

where $f_{\text{token}}(\cdot)$ represents the tokenization process. Subsequently, the token sequence $\mathbf{tok}_i$ is fed into the CLIP encoder to generate a semantic

embedding:

$$p_i = f_{\text{clip}}(\mathbf{tok}_i). \tag{5.2}$$

with $f_{\text{clip}}(\cdot)$ denoting the encoding function. The set of all resulting prompt embeddings is denoted as $\mathcal{P} = \{p_i\}_{i=1}^{N}$.

These embeddings capture high-level contextual information from the textual prompts and are used as additional conditioning inputs in our diffusion model, thereby enhancing the reconstruction performance across different dose levels.

Multi-modal Diffusion Reconstruction: Traditional diffusion models generate images through a two-stage process: a forward diffusion process that gradually adds noise to the input data, and a reverse denoising process that reconstructs the original image. In our multi-modal extension, additional conditioning information—such as textual prompts or clinical parameters—is incorporated into the reverse process to guide the reconstruction. We have defined the traditional diffusion process (along with the definitions for LPET images, SPET images, and spatial consistency guidance) in Chapter 4, and the diffusion process that integrates multi-modal conditions in Chapter 3.

Here, we extend the conventional formulation by integrating prompt embeddings p_i into the reconstruction process. The resulting multi-modal diffusion reconstruction is formulated as:

$$\hat{s}_i = \hat{D}(\mathbf{x}_i^G \mid \mathbf{x}_i^L, p_i). \tag{5.3}$$

where $\hat{s}_i$ denotes the reconstructed SPET image for the ith slice, $\mathbf{x}_i^L$ is the corresponding LPET image, and p_i represents the encoded prompt guidance. This formulation demonstrates how the reverse denoising process leverages both the low-dose image and the supplementary semantic information from the prompt embeddings, ultimately enhancing reconstruction performance across varying dose levels.

Together, these components form the basis of our multimodal diffusion reconstruction process, providing the necessary background for the detailed description of the CWD-PET framework in subsequent sections.

5.2.2 Overview

An overview of the proposed CWD-PET framework is illustrated in Figure 5.2. Building upon our previous WiD-PET model, CWD-PET extends its capabilities by incorporating multi-modal input information and enabling cross-dose level reconstruction. This enhancement allows our framework to robustly handle LPET images acquired at various dose levels while integrating additional guidance from prompt information.

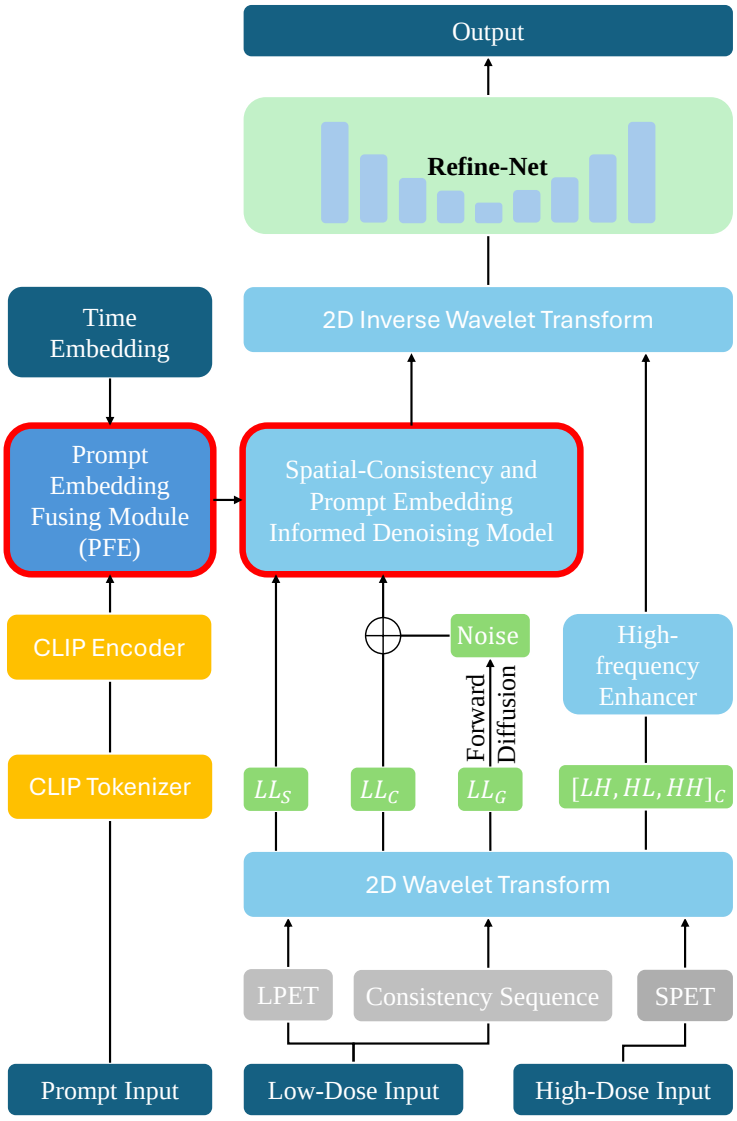


Figure 5.2: Overview of WCD-PET framework, the structure of PEF and Denoising model is in fig 5.4 and fig 5.3.

Specifically, the framework accepts inputs that contain the SPET images, the LPET condition, dose-specific prompt guidance, and spatial sequence guidance derived from adjacent slices. Like mentioned in Chapter 4 Initially, these inputs are decomposed via a Haar 2D Wavelet Transform (2D-DWT) module. This decomposition splits each input into 2 components, one is the low-frequency component, $\mathbf{LL}_x$, and the other is high-frequency components, $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_x$. After that, the low-frequency component $\mathbf{LL}_g$ will undergo a forward diffusion process before being merged with the condition component $\mathbf{LL}_c$ and fed into the denoising model. Additionally, $\mathbf{LL}_S$ from the spatial sequence is employed to provide rich spatial consistency guidance via the SCFE and SCA modules.

A key novelty in CWD-PET is the incorporation of dose prompt guidance. The raw prompt information is first processed by the CLIP tokenizer and then encoded by the CLIP encoder to produce a set of semantic embeddings, denoted as

$$\mathcal{P} = \{p_i\}_{i=1}^{N}.$$

These embeddings are subsequently fused with the time embeddings in a dedicated Prompt Embedding Fusion (PEF) module. This integration effectively injects dose-level context into the reverse diffusion process, enhancing the model's ability to generalize across different LPET dose levels.

Subsequently, the high-frequency components $[\mathbf{LH}, \mathbf{HL}, \mathbf{HH}]_x$ are refined using a High-Frequency Enhancer (HFE) to restore fine details and textures. The refined high-frequency features, together with the reconstructed low-frequency components, are then fused through an inverse 2D Haar Wavelet Transform (2D-IDWT) to generate an intermediate reconstruction. This intermediate result is further processed by a lightweight Refinement-Net (RFN) specifically designed to enhance overall image quality, yielding the final reconstructed image.

This framework, by integrating multi-modal prompt guidance and

novel fusion mechanisms, effectively addresses the challenges associated with cross-dose PET reconstruction, marking a significant advancement over our previous work.

5.2.3 Cross-dose Refine Fast Wavelet-informed Diffusion Architecture

To address the more advanced cross-dose LPET reconstruction task, we extend the fast wavelet-informed diffusion architecture to meet the demands of multi-dose reconstruction—where additional guidance information is required and overall reconstruction quality must be further enhanced. Specifically, we propose an improved spatial-consistency and prompt embedding informed denoising model that incorporates a Prompt Embedding Fusion (PEF) mechanism, as well as a Refinement-Net (RFN) that further boosts reconstruction quality after the IDWT operation.

It is important to note that our overall framework continues to leverage the advantages of the fast wavelet-informed diffusion architecture, including the DWT-IDWT module, the High-Frequency Enhancer (HFE), and the spatial guidance mechanisms—namely, the Spatial-Consistency Feature Extractor (SCFE) and Spatial-Consistency Attention (SCA) modules—used within the denoising model. The specific implementations of these modules have already been detailed in Chapter 4; therefore, in this chapter, we focus solely on introducing the newly added modules and mechanisms.

Spatial-consistency and Prompt embedding Informed Denoising Model

The denoising model is designed to reconstruct noise-free images from the inputs of the low-frequency components (illustrated in Figure 5.3). It adopts a self-attention U-Net architecture enhanced with spatial consistency guidance and cross-modality guidance (fused embedding guidance provided by the PEF module) for more effective feature encoding. The overall architecture comprises L encoder layers, a bottleneck layer, and L decoder layers for upsampling. In contrast to the denoising model

used in Chapter 4, our denoising model incorporates a novel Prompt Embedding Fusing module (PEF).

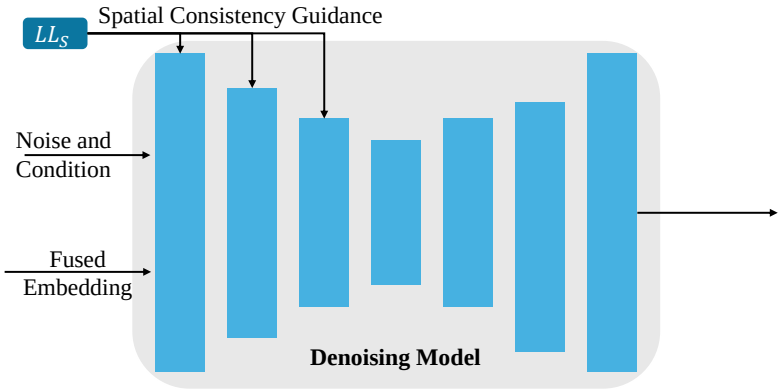


Figure 5.3: Spatial-Consistency and Prompt Embedding Informed Denoising Model

Prompt Embedding Fusing Module (PEF) accepts two inputs: the temporal embedding generated from the diffusion time step, denoted as e_t, and the prompt embedding p_i extracted from the CLIP encoder. Notably, the prompt embedding p_i encodes LPET dose information as a cross-modality prompt, setting it apart from the predominantly image-based inputs in our framework. The structure of PEF is illustrated in Figure 5.4

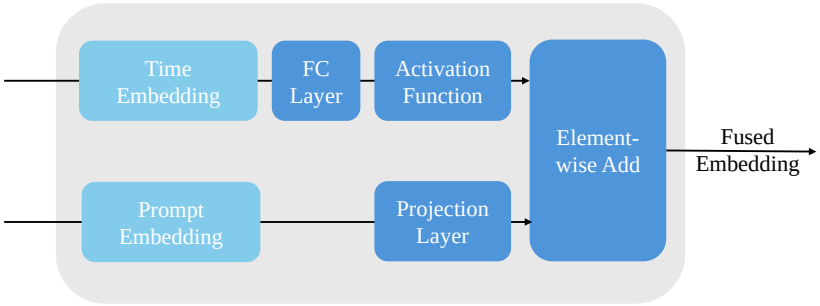


Figure 5.4: The structure of Prompt Embedding Fusing (PEF) module.

Within the PEF, the temporal embedding e_t is first processed through a fully connected layer followed by a nonlinear activation function, and thereby capturing the dynamic characteristics inherent to the diffusion process. Concurrently, the prompt embedding p_i, which encapsulates high-level semantic information, is aggregated via mean pooling across

its tokens, which can effectively condensing multiple token representations into a single vector. To ensure compatibility, the aggregated prompt embedding is then transformed via a projection layer, aligning its dimensionality with that of e_t.

Finally, the transformed prompt embedding and e_t are fused through element-wise addition to yield a unified conditional embedding. This fused embedding leverages the advantages of multimodal integration by providing enhanced semantic guidance and improved robustness in capturing cross-modality features, ultimately contributing to superior image reconstruction quality.

In our enhanced denoising model, the fused conditional embedding produced by the Prompt Embedding Fusing (PEF) module is directly integrated into the network's intermediate layers. This fused embedding provides rich, cross-modal guidance that informs the reverse diffusion process by incorporating LPET dose information alongside the temporal dynamics captured by e_t. In parallel, the spatial consistency modules (SCFE and SCA), as introduced in Chapter 4, continue to supply robust spatial guidance, ensuring that the reconstructed images maintain coherent structure and continuity across slices. Together, these components enable our model to effectively address the challenges associated with cross-dose LPET reconstruction.

Refinement-Net (RFN)

Due to the increased complexity of cross-dose reconstruction—where the differences between input data are more pronounced—a further refinement step is essential to achieve robust and high-quality reconstruction. To this end, and to ensure that the overall reconstruction quality is maximized without incurring significant additional computational costs, we introduce a lightweight Refinement-Net (RFN) that operates after the inverse 2D Haar Wavelet Transform (2D-IDWT).

Built on a U-Net style architecture, the RFN comprises a series of

encoding and decoding layers that include downsampling and upsampling operations, enabling the capture of both global contextual information and fine local details. Skip connections are employed throughout the network to preserve spatial information and facilitate effective multi-scale feature fusion.

Despite its efficacy in enhancing reconstruction quality—smoothing residual artifacts and recovering subtle structures—the RFN is designed to be computationally efficient. Its lightweight structure ensures that it adds minimal overhead to the overall model, thereby preserving the high efficiency of our framework. This balance between improved quality and low computational cost is critical for handling the complexities associated with cross-dose LPET reconstruction, where maintaining fast inference is essential.

5.2.4 Loss Function

We adopt a comprehensive loss function similar to that used in our previous work, with the only modification being that, due to the introduction of the Refinement-Net, the loss is computed on the refined reconstruction output. For brevity, we present the same formulations below.

The noise loss, $\mathcal{L}_{\text{noise}}$, minimizes the discrepancy between the Gaussian noise $\mathbf{n}$ added during the forward diffusion process and the predicted noise $\hat{\mathbf{n}}$ output by the denoising model:

$$\mathcal{L}_{\text{noise}} = \|\mathbf{n} - \hat{\mathbf{n}}\|_2^2. \tag{5.4}$$

To enhance fine details and preserve high-frequency information, we introduce a high-frequency loss, $\mathcal{L}_{\text{hf}}$, which combines an ℓ_2 loss for the high-frequency components with total variation (TV) regularization:

$$\mathcal{L}_{\text{hf}} = \lambda_{\ell_2} \|\mathbf{x}_{\text{hf}} - \hat{\mathbf{x}}_{\text{hf}}\|_2^2 + \lambda_{\text{tv}} \, \text{TV}(\hat{\mathbf{x}}_{\text{hf}}). \tag{5.5}$$

where λ_{ℓ_2} and λ_{tv} are weighting factors balancing the ℓ_2 loss and TV regularization. Here, $\mathbf{x}_{\text{hf}}$ denotes the concatenation of the high-frequency

components of the SPET image (serving as ground truth), and $\hat{\mathbf{x}}_{\text{hf}}$ represents the corresponding high-frequency output produced by our model.

Additionally, we define a reconstruction loss, $\mathcal{L}_{\text{img}}$, using an ℓ_1 loss to measure the difference between the original SPET image $\mathbf{x}^G$ and the reconstructed image $\hat{\mathbf{x}}^G$ (i.e., the output from the Refinement-Net after the IDWT operation):

$$\mathcal{L}_{\text{img}} = \|\mathbf{x}^G - \hat{\mathbf{x}}^G\|_1. \tag{5.6}$$

Combining these components, the final loss function is formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{noise}} + \mathcal{L}_{\text{hf}} + \mathcal{L}_{\text{img}}. \tag{5.7}$$

5.3 Experiments and Results

In this section, we first detail the implementation of our proposed CWD-PET method, and then, we evaluate the performance of our method on the UDPET dataset [59], which includes LPET images acquired at multiple low dose levels. Comprehensive experiments are conducted to assess the effectiveness of our novel framework, including detailed ablation studies. Both qualitative and quantitative results are presented to validate the performance of our approach.

5.3.1 Implementation Details

Our proposed method leverages cross-dose LPET images acquired at $1/2$, $1/4$, $1/20$, $1/50$, and $1/100$ dose levels, with corresponding normal-dose SPET images (i.e., $1/1$ dose) serving as the ground truth. Both LPET and SPET images are resized to 128×128 for input. The framework is implemented in PyTorch and executed on an NVIDIA RTX 3090 GPU. During training, data augmentation techniques—such as random cropping and flipping—are applied to the input images. The initial learning rate is set to 1×10^{-4} and decays by a factor of 0.8 every 5000 iterations. Furthermore, the 2D-DWT-IDWT module is configured with a transformation scale of 2.

5.3.2 Evaluation Metrics

In this section, we briefly summarize the evaluation metrics used to assess the reconstruction quality of our method. As detailed in previous chapters, we still employ PSNR, SSIM, and NMSE for quantitative evaluation, where higher PSNR and SSIM values and lower NMSE indicate better performance. Inference time is measured as the average duration required to generate a full 3D SPET image from whole-brain LPET slices. For brevity, we refer readers to Chapter 3 for a comprehensive description of these metrics.

5.3.3 Experimental Results

The quantitative performance of our proposed method, alongside various baseline models, is summarized in Table 5.1 and Table 5.2. In our evaluation, we compare our approach with the general diffusion-based model DDPM [32], the GAN-based model Still-GAN [47], as well as state-of-the-art PET reconstruction methods including CDM-GAN [29] and PET-Unet [11]. These baselines were chosen primarily because their publicly available code facilitates reproducibility.

It is important to note that these baseline models are optimized for single-dose reconstruction. Since the cross-dose reconstruction problem has not been extensively explored in prior work, we have selected SOTA single-dose models as the comparison benchmark. If our cross-dose reconstruction results, particularly in the ultra-low-dose regime, consistently outperform these models, it demonstrates the overall superiority of our approach in handling diverse dose levels.

Table 5.1 presents the comparison results on ultra-low-dose data (with our method trained and tested in a cross-dose setting), while Table 5.2 reports the complete performance of our cross-dose reconstruction model.

Overall Quality As shown in Table 5.1, the quantitative evaluation reveals that although all baseline methods significantly enhance LPET image quality, our proposed CWD-PET consistently outperforms them across 1/20, 1/50, and 1/100 dose levels. It is noteworthy that while

Table 5.1: Comparison of reconstruction at 1/20, 1/50, and 1/100 dose levels (ultra-low dose regime) on the UDPET dataset. PSNR and SSIM are reported in the table, and NMSE is scaled by $\times 10^{-4}$.

Methods	1/100 dose			1/50 dose			1/20 dose			Inference Time
	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	(s/128 slices)
Low dose[59]	15.46	0.46	22.02	21.70	0.70	22.73	22.72	0.77	4.02	–
2D-DDPM[32]	22.68	0.76	7.27	25.86	0.86	5.62	26.11	0.92	3.27	268.00
3D-DDPM(cWDM)[18]	25.16	0.85	14.10	27.40	0.90	4.20	28.81	0.93	6.10	108.05
Still-GAN[47]	23.61	0.83	9.69	24.48	0.85	7.04	25.63	0.90	3.96	26.12
CDM-GAN[29]	23.95	0.84	12.62	26.80	0.83	3.80	28.84	0.91	3.12	25.24
Pet-Unet[11]	23.30	0.81	8.43	24.33	0.84	6.29	27.93	0.91	2.60	**2.02**
CWD-PET(Ours)	**26.98**	**0.89**	**3.60**	**27.49**	**0.91**	**3.10**	**29.95**	**0.92**	**1.70**	21.25

Table 5.2: Full quantitative results of CWD-PET on the UDPET dataset. PSNR and SSIM are reported in the table, and NMSE is scaled by $\times 10^{-4}$.

Methods	1/100 dose			1/50 dose			1/20 dose			1/10 dose			1/4 dose			1/2 dose		
	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE
Low dose[59]	15.46	0.46	22.02	21.70	0.70	22.73	22.72	0.77	4.02	24.32	0.82	2.60	26.36	0.85	0.89	28.50	0.89	1.80
CWD-PET (Ours)	**26.98**	**0.89**	**3.60**	**27.49**	**0.91**	**3.10**	**29.95**	**0.92**	**1.70**	**30.35**	**0.93**	**0.90**	**31.57**	**0.93**	**0.70**	**32.80**	**0.95**	**0.50**

the baseline models—such as CDM-GAN, PET-Unet, DDPM, and Still-GAN—are individually optimized for single-dose reconstruction, our cross-dose model is designed to generalize effectively across a range of dose levels. For instance, at the 1/100 dose level, the LPET images exhibit PSNR, SSIM, and NMSE values of 15.46, 0.46, and 22.02, respectively; in contrast, CWD-PET improves these metrics to 26.98, 0.89, and 3.60. This represents a substantial enhancement over the second-best performer, 3D-DDPM(cWDM), which achieves 25.16, 0.85, and 14.10.

Furthermore, while the general diffusion-based model 2D- and 3D-DDPM outperforms the GAN-based StillGAN, both fall short compared to the specialized single-dose reconstruction models. In contrast, our CWD-PET framework, with its cross-dose design, not only delivers superior performance in terms of high-frequency detail restoration and spatial consistency but also demonstrates robust generalizability across varying dose levels. These results underscore the efficacy of our approach in tackling the complexities of cross-dose LPET reconstruction, setting a new benchmark even against methods optimized solely for

single-dose scenarios.

Visual Comparison A visualization comparison of the 2D methods' reconstruction results is given in Figure 5.5. As can be seen, the comparison results in the ultra-low dose regime clearly demonstrate that the images reconstructed by our CWD-PET model are sharper and exhibit richer details that closely resemble the ground truth. Our method's advantages become even more pronounced at ultra-low dose levels. Moreover, the results presented in (b) further confirm that our cross-dose reconstruction approach delivers effective and high-quality reconstructions across various dose levels. This qualitative evidence is consistent with our quantitative analysis, thereby verifying the effectiveness of our proposed strategies.

Inference Speed Beyond its superior reconstruction performance, our CWD-PET model demonstrates remarkable efficiency during inference. As shown in Table 5.1, the integration of the wavelet transform reduces the inference time by nearly 10-fold—from 268 seconds per 128 slices for 2D-DDPM to just 21.25 seconds per 128 slices (also 5 times less than 3D-DDPM with wavelet decompose transformation module). This inference speed outperforms that of GAN-based models and is second only to Pet-Unet, which, despite its faster inference, produces significantly inferior reconstruction quality. Moreover, with the addition of the Prompt Embedding Fusion (PEF) module and our lightweight Refinement-Net (RFN), our framework not only enhances reconstruction quality but also maintains its fast inference speed. These results underscore the efficiency and effectiveness of our proposed approach for cross-dose PET reconstruction.

5.3.4 Ablation Study

We conducted ablation studies to evaluate the effectiveness of each component in our proposed method, and the results are summarized in Table 4.2. As shown in Row 2, when employing the base model with the DWT-IDWT module and prompt embedding guidance (PEF) but without the high-frequency loss term $\mathcal{L}_{\text{hf}}$ for cross-dose reconstruction, the

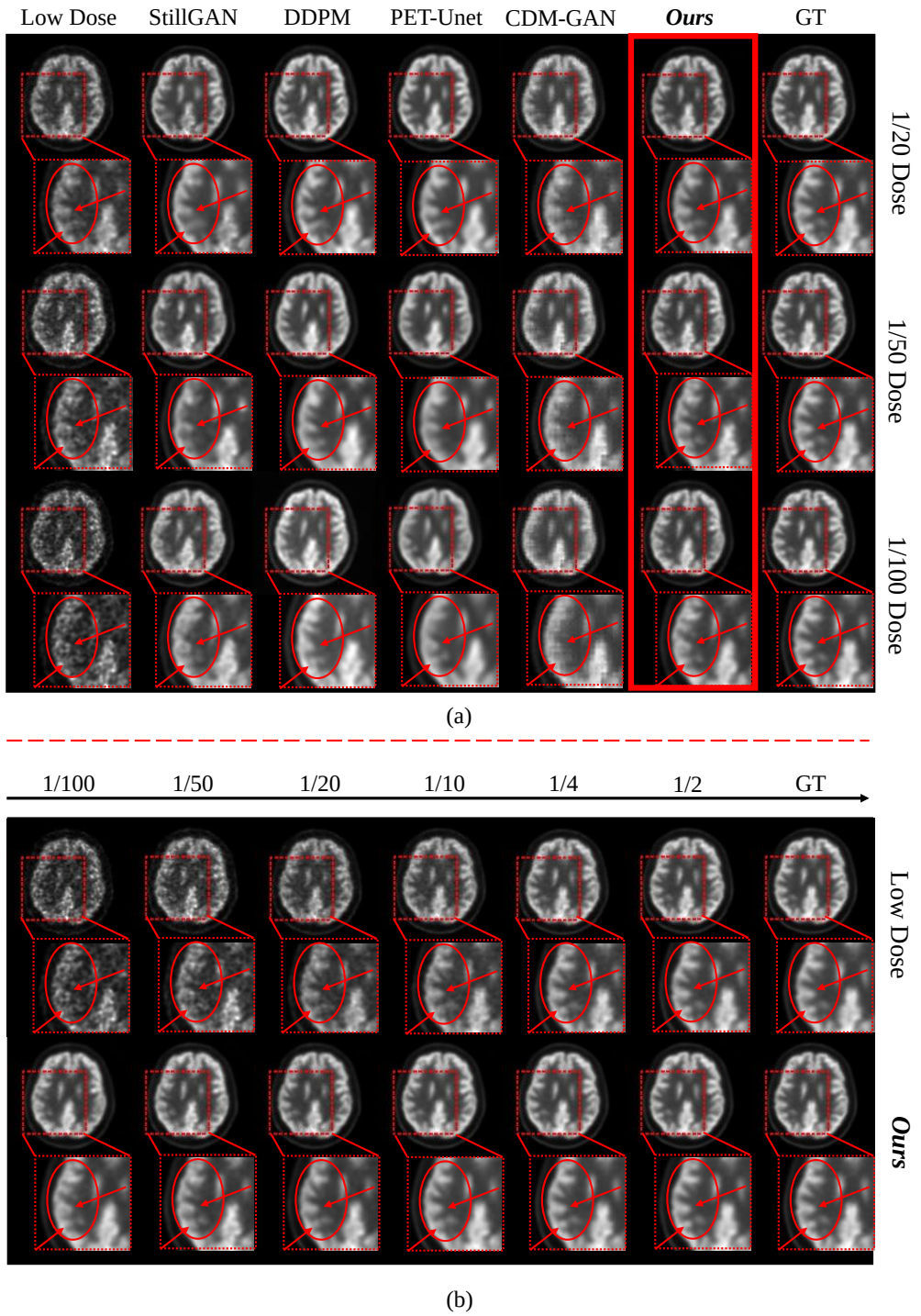


Figure 5.5: Qualitative Reconstruction Results of 2D methods. (a) Visual comparison of reconstructions across different dose levels: the top row shows the full-slice view for each dose level, while the bottom row provides a zoomed-in view of regions highlighted by red boxes in the top row. (b) Reconstruction results of our cross-dose method across various dose levels: in each column, the top row displays the corresponding low-dose PET image (with the last column representing the ground truth), and the bottom row presents the reconstructed image, with dose levels indicated above each column. These views clearly illustrate the enhanced detail recovery and spatial consistency achieved by our approach.

PSNR/SSIM values at dose levels of 1/100, 1/50, 1/20, 1/10, 1/4, and 1/2 are 26.26/0.86, 27.08/0.88, 27.99/0.89, 28.81/0.90, 29.66/0.92, and 30.43/0.93, respectively. When the high-frequency loss term is added (Row 3), the performance improves to 26.55/0.87, 27.42/0.90, 28.51/0.91, 29.53/0.92, 30.78/0.93, and 32.17/0.94. Furthermore, the incorporation of the Spatial-Consistency Feature Extractor (SCFE) in Row 4 yields results of 26.63/0.87, 27.45/0.90, 28.59/0.91, 29.89/0.93, 30.89/0.93, and 32.25/0.94. When the Spatial-Consistency Attention (SCA) module is further combined with SCFE (Row 6), the metrics improve further to 26.91/0.88, 27.95/0.90, 28.89/0.92, 30.09/0.92, 31.07/0.93, and 32.35/0.93. Notably, when using spatial consistency guidance (SCFE and SCA) without the high-frequency loss term (Row 5), the performance is 26.50/0.87, 27.33/0.91, 28.56/0.92, 29.62/0.94, 31.09/0.95, and 32.34/0.94. Finally, with the incorporation of the Refinement-Net (RFN), our full method achieves its best results, with PSNR/SSIM values of 26.98/0.89, 27.49/0.91, 29.95/0.92, 30.35/0.93, 31.57/0.93, and 32.80/0.95 at the corresponding dose levels.

These results demonstrate that integrating all the components yields the highest reconstruction performance in the cross-dose PET reconstruction task, thereby confirming that each module—ranging from the high-frequency loss term to the spatial consistency and refinement mechanisms—significantly contributes to the overall effectiveness of our approach.

5.4 Summary

In this chapter, we extend our previously proposed fast wavelet-informed diffusion model for PET reconstruction (WiD-PET) to address the more advanced and clinically valuable task of cross-dose PET reconstruction. Building upon the wavelet-informed diffusion architecture, we introduce multi-modal guidance in the form of prompt information. By leveraging the CLIP tokenizer and encoder, we obtain prompt embeddings that are effectively fused through a dedicated Prompt Embedding Fusing (PEF) module, thereby enabling the prompt information to robustly

Table 5.3: Results of ablation experiments on cross-dose levels from the UDPET dataset. Columns C1–C7 denote the following components:**DDPM Base Model**, **Wavelet Transformation and HFE**, **SCFE**, **SCA**, **HF loss**, **PEF**, and **RFN**. PSNR and SSIM are reported in the table, and NMSE is scaled by $\times 10^{-4}$.

Components							1/100 dose			1/50 dose			1/20 dose			1/10 dose			1/4 dose			1/2 dose		
C1	C2	C3	C4	C5	C6	C7	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE	PSNR	SSIM	NMSE
✓	✓	·	·	·	✓	·	26.26	0.86	4.40	27.08	0.88	3.70	27.99	0.89	2.90	28.81	0.90	2.50	29.66	0.92	2.20	30.43	0.93	1.90
✓	✓	·	·	✓	✓	·	26.55	0.87	4.10	27.42	0.90	3.30	28.51	0.91	2.70	29.53	0.92	2.30	30.78	0.93	1.90	32.17	0.94	1.60
✓	✓	✓	·	✓	✓	·	26.63	0.87	4.10	27.45	0.90	3.30	28.59	0.91	2.40	29.89	0.93	2.00	30.89	0.93	1.70	32.25	0.94	1.40
✓	✓	✓	✓	·	✓	·	26.50	0.87	4.32	27.33	0.91	3.71	28.56	0.92	2.11	29.62	0.94	1.73	31.09	0.95	1.51	32.34	0.94	1.17
✓	✓	✓	✓	✓	✓	·	26.91	0.88	3.60	27.95	0.90	3.40	28.89	0.92	2.00	30.09	0.92	1.20	31.07	0.93	0.90	32.35	0.94	0.50
✓	✓	✓	✓	✓	✓	✓	**26.98**	**0.89**	**3.60**	**27.49**	**0.91**	**3.10**	**29.95**	**0.92**	**1.72**	**30.35**	**0.93**	**0.90**	**31.57**	**0.93**	**0.70**	**32.8**	**0.95**	**0.50**

guide the reconstruction process. In addition, we propose a Refinement-Network (RFN) to further enhance the reconstruction quality after the inverse Wavelet Transform (IDWT), specifically improving the cross-dose reconstruction results.

We conducted extensive experiments on the UDPET dataset, evaluating our method on LPET images acquired at various dose levels. The experimental results demonstrate that our cross-dose reconstruction model not only exhibits high efficiency but also achieves superior reconstruction quality. In particular, our method outperforms state-of-the-art single-dose optimized models in the ultra-low dose regime, and consistently delivers excellent results across all dose levels. These findings verify the effectiveness and clinical potential of our proposed approach.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

PET imaging plays a crucial role in clinical diagnostics and treatment planning, offering invaluable insights into human metabolism. While leveraging lower doses of radiotracers helps to mitigate potential risks, ensuring high-quality PET reconstruction remains of paramount importance. Over the years, numerous reconstruction approaches have been proposed, beginning with traditional methods and evolving through the deep learning (DL) era—where U-Net, GAN, and related models have markedly improved reconstruction quality. More recently, approaches based on diffusion models have emerged as a promising alternative in image reconstruction, particularly in the medical imaging domain, due to their inherent advantages such as stable training and high-fidelity output.

Despite the advances achieved by traditional, Deep Learning-based, and diffusion-based methods, several challenges persist. These include the lack of spatial consistency in 2D reconstructions, insufficient recovery of fine details, slow inference speeds, and the neglect of cross-dose reconstruction scenarios. In response, this thesis has focused on addressing these critical issues.

Our contributions can be summarized as follows:

- We proposed a novel fast wavelet-informed diffusion model, termed **WiD-PET**, which leverages 2D-DWT to decompose PET images into low- and high-frequency components. The low-frequency parts

are reconstructed via a diffusion-based denoising model guided by spatial consistency modules (SCFE and SCA), while a specially designed High-Frequency Enhancer (HFE) refines the high-frequency components. These components are then reassembled using 2D-IDWT to produce high-quality reconstructions with significantly reduced inference time.

- Building upon WiD-PET, we further introduced the cross-dose PET reconstruction model, the **CWD-PET** framework, to tackle the advanced task of cross-dose PET reconstruction. This extension incorporates multi-modal guidance by integrating prompt information—extracted using CLIP's tokenizer and encoder—and fusing it through a novel Prompt Embedding Fusion (PEF) module within the denoising model. Moreover, a lightweight Refinement Network (RFN) is applied post-IDWT to further enhance reconstruction quality, ensuring robust performance across a wide range of LPET dose levels.
- Extensive experiments on a low-dose PET dataset validate the effectiveness and efficiency of our approach. Our results demonstrate that, even when compared to state-of-the-art models optimized for single-dose reconstruction, our cross-dose model achieves superior reconstruction quality, particularly in the regimes of ultra-low dose, while maintaining a high level of efficiency.

Overall, this thesis establishes a new benchmark for PET image reconstruction by effectively addressing key challenges and paving the way for further advancements in cross-dose and multimodal reconstruction techniques.

6.2 Future Work

There are two promising research directions for future work.

The first is to incrementally improve cross-dose reconstruction. In this thesis, we have explored the cross-dose PET reconstruction scenario by incorporating prompt guidance to provide additional information. In

practice, different LPET dose levels exhibit inherent correlations; therefore, establishing a reconstruction process that begins with the lowest dose level and progressively transitions to higher or multiple dose levels—eventually recovering the SPET image—could better exploit these inter-dose relationships. Moreover, integrating the intrinsic properties of diffusion models with dose-level information (e.g., using dose levels as guidance in the iterative denoising process) may further enhance reconstruction performance. This progressive or staged enhancement of cross-dose reconstruction remains an underexplored yet clinically valuable area.

The second direction focuses on achieving high-efficiency full 3D diffusion-based PET reconstruction. Current diffusion networks require many denoising steps, and when applied to high-resolution 3D data, they incur substantial computational costs and significantly prolong inference times. Addressing this efficiency challenge is critical; future work may involve designing architectures that reduce the dimensionality of 3D inputs, developing robust low-dimensional representations for high-dimensional image features, or inventing novel diffusion strategies that accelerate the denoising process. Such advancements could make 3D diffusion-based PET reconstruction not only feasible but also highly efficient for clinical applications.

In conclusion, given the crucial role of high-quality SPET images in clinical practice and the imperative to reduce patients' exposure to radiotracers, the task of reconstructing high-quality SPET images from LPET inputs will remain a focal research area. We believe that, with the continued evolution of new algorithms, models, and techniques, this field will continue to progress robustly, leading to the development of even more effective reconstruction methods and inspiring further research directions.

6.3 Broader Potential of Image Denoising Techniques

Image denoising techniques, including classical methods as well as advanced deep learning approaches such as diffusion models, aim to enhance image quality by mitigating noise introduced through environmental or technical factors. While this thesis primarily focuses on PET image reconstruction, these denoising methodologies have broad applicability across various domains, including other medical imaging modalities, specialized industrial imaging, general natural image processing, and diverse media formats.

Looking ahead, there is significant potential to extend and adapt denoising algorithms beyond medical imaging. For example, in natural image processing, denoising techniques have been successfully applied to enhance visual quality in photography, videography, and broadcasting media. Algorithm-driven denoising devices, particularly those leveraging deep learning, offer promising directions for portable and cost-effective image enhancement solutions. Furthermore, there is a growing demand for refined denoising methods that align with human perceptual needs in entertainment and multimedia industries, where improving image quality in a visually coherent manner is critical.

Overall, the continuous advancement of denoising techniques, especially diffusion-based methods capable of not only suppressing noise but also generating realistic image details through a learned denoising process, holds great promise for improving both clinical imaging outcomes and a wide array of visual media applications. These developments are expected to drive further research and innovation in image restoration and synthesis.

Bibliography

[1] A. M. Alessio, C. W. Stearns, S. Tong, S. G. Ross, S. Kohlmyer, A. Ganin, and P. E. Kinahan. "Application and evaluation of a measured spatially variant system model for PET image reconstruction". In: *IEEE transactions on medical imaging* 29.3 (2010), pp. 938–949.

[2] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan. "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions". In: *Journal of Big Data* 8 (2021), pp. 1–74. URL: https://link.springer.com/article/10.1186/s40537-021-00444-8.

[3] S. Anand, H. Singh, and A. Dash. "Clinical applications of PET and PET-CT". In: *Medical Journal Armed Forces India* 65.4 (2009), pp. 353–358.

[4] J. Baek, V. F. Farias, A. Georgescu, R. Levi, T. Peng, D. Sinha, J. Wilde, and A. Zheng. "The limits to learning a diffusion model". In: *Proceedings of the 22nd ACM Conference on Economics and Computation*. 2021, pp. 130–131.

[5] D. L. Bailey, D. W. Townsend, P. E. Valk, and M. N. Maisey, eds. *Positron Emission Tomography: Basic Sciences*. 1st ed. Springer London, 2003. ISBN: 978-1-85233-485-7. DOI: 10.1007/b136169. URL: https://link.springer.com/book/10.1007/b136169.

[6] N. A. Benatar, B. F. Cronin, and M. J. O'doherty. "Radiation Dose Rates from Patients Undergoing Positron Emission Tomography: Implications for Technologists and Waiting Areas". In: *European Journal of Nuclear Medicine* 27 (2000), pp. 583–589. DOI: 10.1007/s002590050546. URL: https://doi.org/10.1007/s002590050546.

[7] A. Buades, B. Coll, and J.-M. Morel. "A non-local algorithm for image denoising". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 2. IEEE. 2005, pp. 60–65. DOI: 10.1109/CVPR.2005.38.

[8] A. Çayir, I. Yenidoğan, and H. Dağ. "Feature Extraction Based on Deep Learning for Some Traditional Machine Learning Methods". In: *2018 3rd International Conference on Computer Science and Engineering (UBMK)*. 2018, pp. 494–497. DOI: 10.1109/UBMK.2018.8566383.

[9] C. Chan, R. Fulton, D. D. Feng, and S. Meikle. "Median non-local means filtering for low SNR image denoising: Application to PET with anatomical knowledge". In: *IEEE Nuclear Science Symposuim & Medical Imaging Conference*. IEEE. 2010, pp. 3613–3618.

[10] H. Chen, X. Wang, Y. Zhou, B. Huang, Y. Zhang, W. Feng, H. Chen, Z. Zhang, S. Tang, and W. Zhu. *Multi-Modal Generative AI: Multimodal LLM, Diffusion and Beyond*. 2024. arXiv: 2409.14993 [cs.AI]. URL: https://arxiv.org/abs/2409.14993.

[11] K. Chen, E. Gong, F. de Carvalho Macruz, J. Xu, A. Boumis, M. Khalighi, K. Poston, S. Sha, M. Greicius, E. Mormino, J. Pauly, S. Srinivas, and G. Zaharchuk. "Ultra-Low-Dose 18F-Florbetaben Amyloid PET Imaging Using Deep Learning with Multi-Contrast MRI Inputs". In: *Radiology* 290.3 (Mar. 2019). Epub 2018 Dec 11. Erratum in: Radiology. 2020 Sep;296(3):E195. doi: 10.1148/radiol.2020202527, pp. 649–656. DOI: 10.1148/radiol.2018180940.

[12] M. Coşkun, Ö. Yıldırım, A. Uçar, and Y. Demır. "An Overview of Popular Deep Learning Methods". In: *European Journal of Technique (EJT)* 7.2 (2017), pp. 165–176. URL: https://dergipark.org.tr/en/pub/ejt/issue/34562/403498.

[13] J. Cui, Y. Wang, L. Wen, P. Zeng, X. Wu, J. Zhou, and D. Shen. "Image2Points: A 3D Point-Based Context Clusters GAN for High-Quality Pet Image Reconstruction". In: *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2024, pp. 1726–1730. DOI: 10.1109/ICASSP48485.2024.10446360.

[14] J. Cui, X. Zeng, P. Zeng, B. Liu, X. Wu, J. Zhou, and Y. Wang. "MCAD: Multi-modal Conditioned Adversarial Diffusion Model for High-Quality PET Image Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*. Ed. by M. G. Linguraru, Q. Dou, A. Feragen, S. Giannarou, B. Glocker, K. Lekadir, and J. A. Schnabel. Cham: Springer Nature Switzerland, 2024, pp. 467–477. ISBN: 978-3-031-72104-5. URL: https://arxiv.org/abs/2406.13150.

[15] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. "Image denoising by sparse 3-D transform-domain collaborative filtering". In: *IEEE Transactions on Image Processing* 16.8 (2007), pp. 2080–2095. DOI: 10.1109/TIP.2007.901238.

[16] D. Delbeke and W. H. Martin. "Positron Emission Tomography Imaging in Oncology". In: *Radiologic Clinics of North America* 39.5 (2001), pp. 883–917. ISSN: 0033-8389. DOI: 10.1016/S0033-8389(05)70319-5. URL: https://www.sciencedirect.com/science/article/pii/S0033838905703195.

[17] Y. Fan, H. Liao, S. Huang, Y. Luo, H. Fu, and H. Qi. "A survey of emerging applications of diffusion probabilistic models in MRI". In: *Meta-Radiology* (2024), p. 100082.

[18] P. Friedrich, A. Durrer, J. Wolleb, and P. C. Cattin. *cWDM: Conditional Wavelet Diffusion Models for Cross-Modality 3D Medical Image Synthesis*. 2024. arXiv: 2411.17203 [eess.IV]. URL: https://arxiv.org/abs/2411.17203.

[19] D. Ganguly, S. Chakraborty, M. Balitanas, and T.-h. Kim. "Medical imaging: A review". In: *International Conference on Security-Enriched Urban Computing and Smart Grid*. Springer. 2010, pp. 504–516.

[20] K. K. Ghosh, P. Padmanabhan, C.-T. Yang, S. Mishra, C. Halldin, and B. Gulyás. "Dealing with PET radiometabolites". In: *EJNMMI research* 10 (2020), pp. 1–17.

[21] L. Gondara. "Medical Image Denoising Using Convolutional Denoising Autoencoders". In: *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. 2016, pp. 241–246. DOI: 10.1109/ICDMW.2016.0041.

[22] K. Gong, J. Guan, K. Kim, X. Zhang, J. Yang, Y. Seo, G. El Fakhri, J. Qi, and Q. Li. "Iterative PET Image Reconstruction Using Convolutional Neural Network Representation". In: *IEEE Transactions on Medical Imaging* 38.3 (2019), pp. 675–685. DOI: 10.1109/TMI.2018.2869871.

[23] K. Gong, K. Johnson, G. El Fakhri, Q. Li, and T. Pan. "PET Image Denoising Based on Denoising Diffusion Probabilistic Model". In: *European Journal of Nuclear Medicine and Molecular Imaging* 51.2 (2024), pp. 358–368. DOI: 10.1007/s00259-023-06417-8. URL: https://doi.org/10.1007/s00259-023-06417-8.

[24] K. Gong, K. Kim, J. Cui, D. Wu, and Q. Li. "The evolution of image reconstruction in PET: from filtered back-projection to artificial intelligence". In: *PET clinics* 16.4 (2021), pp. 533–542.

[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Networks". In: *Communications of the ACM* 63.11 (Oct. 2020), pp. 139–144. ISSN: 0001-0782. DOI: 10.1145/3422622. URL: https://doi.org/10.1145/3422622.

[26] A. Graikos, N. Malkin, N. Jojic, and D. Samaras. "Diffusion models as plug-and-play priors". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 14715–14728.

[27] M. A. Haidekker. *Medical imaging technology*. 2013.

[28] Z. Han, Y. Wang, L. Zhou, P. Wang, B. Yan, J. Zhou, Y. Wang, and D. Shen. "Contrastive Diffusion Model with Auxiliary Guidance for Coarse-to-Fine PET Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Cham: Springer Nature Switzerland, 2023, pp. 239–249. ISBN: 978-3-031-43999-5. URL: https://arxiv.org/abs/2308.10157.

[29] Z. Han, Y. Wang, L. Zhou, P. Wang, B. Yan, J. Zhou, Y. Wang, and D. Shen. *Contrastive Diffusion Model with Auxiliary Guidance for Coarse-to-Fine PET Reconstruction*. 2023. arXiv: 2308.10157 [eess.IV]. URL: https://arxiv.org/abs/2308.10157.

[30] F. He, T. Liu, and D. Tao. "Why ResNet Works? Residuals Generalize". In: *IEEE Transactions on Neural Networks and Learning Systems* 31.12 (2020), pp. 5349–5362. DOI: 10.1109/TNNLS.2020.2966319.

[31] D. Hellwig, N. C. Hellwig, S. Boehner, T. Fuchs, R. Fischer, and D. Schmidt. "Artificial intelligence and deep learning for advancing PET image reconstruction: State-of-the-art and future directions". In: *Nuklearmedizin-NuclearMedicine* 62.06 (2023), pp. 334–342.

[32] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[33] J. Ho, A. Jain, and P. Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 6840–6851. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[34] C. K. Hoh. "Clinical use of FDG PET". In: *Nuclear medicine and biology* 34.7 (2007), pp. 737–742.

[35] A. E. Ilesanmi and T. O. Ilesanmi. "Methods for Image Denoising Using Convolutional Neural Network: A Review". In: *Complex & Intelligent Systems* 7.5 (2021), pp. 2179–2198. URL: https://link.springer.com/article/10.1007/s40747-021-00428-4.

[36] Y. Inoue. "Radiation Dose Modulation of Computed Tomography Component in Positron Emission Tomography/Computed Tomography". In: *Seminars in Nuclear Medicine* 52.2 (2022). Radiation Exposure and Dosimetry, pp. 157–166. ISSN: 0001-2998. DOI: 10.1053/j.semnuclmed.2021.11.009. URL: https://www.sciencedirect.com/science/article/pii/S0001299821000969.

[37] A. Iriarte, R. Marabini, S. Matej, C. O. S. Sorzano, and R. M. Lewitt. "System models for PET statistical iterative reconstruction: A review". In: *Computerized Medical Imaging and Graphics* 48 (2016), pp. 30–48.

[38] H. Jadvar and J. Parker. "PET radiotracers". In: *Clinical PET and PET/CT* (2005), pp. 45–67.

[39] C. Jiang, Y. Pan, M. Liu, L. Ma, X. Zhang, J. Liu, X. Xiong, and D. Shen. "PET-Diffusion: Unsupervised PET Enhancement Based on the Latent Diffusion Model". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*. Ed. by H. Greenspan, A. Madabhushi, P. Mousavi, S. Salcudean, J. Duncan, T. Syeda-Mahmood, and R. Taylor. Cham: Springer Nature Switzerland, 2023, pp. 3–12. ISBN: 978-3-031-43907-0. DOI: https://doi.org/10.1007/978-3-031-43907-0_1.

[40] Z. Jin. "Advancements in Diffusion Models for Image Generation: A Comparative Analysis of DDPM, LDM, and DDIM". In: *Applied and Computational Engineering* 104 (2024), pp. 96–103.

[41] J. Kang, Y. Gao, F. Shi, D. S. Lalush, W. Lin, and D. Shen. "Prediction of standard-dose brain PET image by using MRI and low-dose brain [18F] FDG PET images". In: *Medical Physics* 42.9 (2015), pp. 5301–5309. DOI: 10.1118/1.4928400. URL: https://doi.org/10.1118/1.4928400.

[42] N. A. Karakatsanis, E. Fokou, and C. Tsoumpas. "Dosage Optimization in Positron Emission Tomography: State-of-the-Art Methods and Future Prospects". In: *American Journal of Nuclear Medicine and Molecular Imaging* 5.5 (2015), p. 527. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4620179/.

[43] J. H. Kim, I. J. Ahn, W. H. Nam, and J. B. Ra. "An effective post-filtering framework for 3-D PET image denoising based on noise and sensitivity characteristics". In: *IEEE Transactions on Nuclear Science* 62.1 (2014), pp. 137–147.

[44] Y. Lai. "A Comparison of Traditional Machine Learning and Deep Learning in Image Recognition". In: *Journal of Physics: Conference Series* 1314.1 (Oct. 2019), p. 012148. DOI: 10.1088/1742-6596/1314/1/012148. URL: https://dx.doi.org/10.1088/1742-6596/1314/1/012148.

[45] V. J. Lowe, F. G. Duhaylongsod, E. F. Patz, D. M. Delong, J. M. Hoffman, W. G. Wolfe, and R. E. Coleman. "Pulmonary Abnormalities and PET Data Analysis: A Retrospective Study". In: *Radiology* 202.2 (1997), pp. 435–439. DOI: 10.1148/radiology.202.2.9015070. URL: https://doi.org/10.1148/radiology.202.2.9015070.

[46] Y. Luo, Y. Wang, C. Zu, B. Zhan, X. Wu, J. Zhou, D. Shen, and L. Zhou. "3D Transformer-GAN for High-Quality PET Reconstruction". In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. Ed. by M. de Bruijne, P. C. Cattin, S. Cotin, N. Padoy, S. Speidel, Y. Zheng, and C. Essert. Cham: Springer International Publishing, 2021, pp. 276–285. ISBN: 978-3-030-87231-1.

[47] Y. Ma, J. Liu, Y. Liu, H. Fu, Y. Hu, J. Cheng, H. Qi, Y. Wu, J. Zhang, and Y. Zhao. "Structure and Illumination Constrained GAN for Medical Image Enhancement". In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3955–3967. DOI: 10.1109/TMI.2021.3101937.

[48] G. Muehllehner and J. S. Karp. "Positron Emission Tomography". In: *Physics in Medicine & Biology* 51.13 (2006), R117. URL: https://doi.org/10.1088/0031-9155/51/13/R08.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[50] A. J. Reader and B. Pan. "AI for PET image reconstruction". In: *The British journal of radiology* 96.1150 (2023), p. 20230292.

[51] A. J. Reader and H. Zaidi. "Advances in PET image reconstruction". In: *PET clinics* 2.2 (2007), pp. 173–190.

[52] A. J. Reader and H. Zaidi. "Advances in PET Image Reconstruction". In: *PET Clinics* 2.2 (2007). PET Instrumentation and Quantification, pp. 173–190. ISSN: 1556-8598. DOI: 10.1016/j.cpet.2007.08.001. URL: https://www.sciencedirect.com/science/article/pii/S1556859807000193.

[53] S. Remmele, J. Hesser, H. Paganetti, and T. Bortfeld. "A deconvolution approach for PET-based dose reconstruction in proton radiotherapy". In: *Physics in Medicine & Biology* 56.23 (2011), p. 7601.

[54] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. "Image Super-Resolution via Iterative Refinement". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.4 (2022), pp. 4713–4726.

[55] A. Sanaat, I. Shiri, H. Arabi, I. Mainta, R. Nkoulou, and H. Zaidi. "Deep learning-assisted ultra-fast/low-dose whole-body PET/CT imaging". In: *European Journal of Nuclear Medicine and Molecular Imaging* 48 (2021), pp. 2405–2415. URL: https://doi.org/10.1007/s00259-020-05167-1.

[56] H. Sasaki, C. G. Willcocks, and T. P. Breckon. *UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models*. 2021. arXiv: 2104.05358 [cs.CV]. URL: https://arxiv.org/abs/2104.05358.

[57] D. J. Schlyer. "PET tracers and radiochemistry". In: *Annals-Academy of Medicine Singapore* 33.2 (2004), pp. 146–154.

[58] W. El-Shafai, A. A. Mahmoud, A. M. Ali, E.-S. M. El-Rabaie, T. E. Taha, O. F. Zahran, A. S. El-Fishawy, N. F. Soliman, A. A. Alhussan, and F. E. Abd El-Samie. "Deep CNN Model for Multimodal Medical Image Denoising". In: *Computers, Materials & Continua* 73.2 (2022), pp. 3795–3814. ISSN: 1546-2226. DOI: 10.32604/cmc.2022.029134. URL: http://www.techscience.com/cmc/v73n2/48384.

[59] K. Shi, R. Guo, S. Xue, A. Rominger, and B. Li. *Ultra-low Dose PET Imaging Challenge 2022*. 2022. URL: https://zenodo.org/records/6361846.

[60] A. K. Shukla and U. Kumar. "Positron Emission Tomography: An Overview". In: *Journal of Medical Physics* 31.1 (2006), pp. 13–21. URL: https://doi.org/10.4103/0971-6203.25665.

[61] K. K. Shung, M. Smith, and B. M. Tsui. *Principles of medical imaging*. Academic Press, 2012.

[62] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 2256–2265. URL: https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[63] J. Song, C. Meng, and S. Ermon. "Denoising Diffusion Implicit Models". In: *CoRR* abs/2010.02502 (2020). arXiv: 2010.02502. URL: https://arxiv.org/abs/2010.02502.

[64] B. Tan, Y. Xue, L. Bi, and J. Kim. "Full-TrSUN: A Full-Resolution Transformer UNet for High Quality PET Image Synthesis". In: *Machine Learning in Medical Imaging*. Ed. by X. Xu, Z. Cui, I. Rekik, X. Ouyang, and K. Sun. Cham: Springer Nature Switzerland, 2025, pp. 238–247. ISBN: 978-3-031-73284-3.

[65] G. Tarantola, F. Zito, and P. Gerundini. "PET instrumentation and reconstruction algorithms in whole-body applications". In: *Journal of Nuclear Medicine* 44.5 (2003), pp. 756–769.

[66] G. Wang and J. Qi. "PET Image Reconstruction Using Kernel Method". In: *IEEE Transactions on Medical Imaging* 34.1 (2015), pp. 61–71. DOI: 10.1109/TMI.2014.2343916.

[67] Y. Wang, G. Ma, L. An, F. Shi, P. Zhang, D. S. Lalush, X. Wu, Y. Pu, J. Zhou, and D. Shen. "Semisupervised Tripled Dictionary Learning for Standard-Dose PET Image Prediction Using Low-Dose PET and Multimodal MRI". In: *IEEE Transactions on Biomedical Engineering* 64.3 (2017), pp. 569–579. DOI: 10.1109/TBME.2016.2564440.

[68] T. Xia, A. M. Alessio, B. De Man, R. Manjeshwar, E. Asma, and P. E. Kinahan. "Ultra-low dose CT attenuation correction for PET/CT". In: *Physics in Medicine & Biology* 57.2 (2011), p. 309. DOI: 10.1088/0031-9155/57/2/309. URL: https://doi.org/10.1088/0031-9155/57/2/309.

[69] J. Zhai, S. Zhang, J. Chen, and Q. He. "Autoencoder and Its Various Variants". In: *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2018, pp. 415–419. DOI: 10.1109/SMC.2018.00080.

[70] D. Zhang. *Fundamentals of Image Data Mining: Analysis, Features, Classification and Retrieval*. Springer Nature, 2019. URL: https://link.springer.com/book/10.1007/978-3-030-69251-3.

[71] Z. Zhang, M. Li, and J. Yu. "On the Convergence and Mode Collapse of GAN". In: *SIGGRAPH Asia 2018 Technical Briefs*. 2018, pp. 1–4. DOI: 10.1145/3283254.3283282.

[72] K. Zhao, L. Zhou, S. Gao, X. Wang, Y. Wang, X. Zhao, H. Wang, K. Liu, Y. Zhu, and H. Ye. "Study of low-dose PET image recovery using supervised learning with CycleGAN". In: *PLOS ONE*

15.9 (2020), e0238455. URL: https://pubmed.ncbi.nlm.nih.gov/32886683/.

[73] Z. Zhao, H. Bai, Y. Zhu, J. Zhang, S. Xu, Y. Zhang, K. Zhang, D. Meng, R. Timofte, and L. Van Gool. "DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion". In: 2023. arXiv: 2303.06840 [cs.CV]. URL: https://arxiv.org/abs/2303.06840.

[74] L. Zhou, J. D. Schaefferkoetter, I. W. Tham, G. Huang, and J. Yan. "Supervised learning with CycleGAN for low-dose FDG PET image denoising". In: *Medical Image Analysis* 65 (2020), p. 101770. ISSN: 1361-8415. DOI: 10.1016/j.media.2020.101770. URL: https://www.sciencedirect.com/science/article/pii/S1361841520301341.