Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Health

SBSDM: A Style-aware Bidirectional Stream Diffusion Model for CT-to-PET Synthesis

Jiahao Zheng[1], **Yu Tang**[2✉], **Caiwen Jiang**[3], **Zhanjie Zhang**[4], **Yongcan Luo**[1] and **Dapeng Wu**[1]
[1]City University of Hong Kong
[2]Shanghai University
[3]Mayo Clinic in Arizona
[4]Zhejiang University

Abstract

CT-to-PET synthesis aims to synthesize PET images from the widely available and lower-cost CT scans to address the high cost and additional radiation exposure associated with PET scanning. However, CT-to-PET synthesis faces two key challenges due to the sequential correlation of volumetric imaging: preserving smooth transition between adjacent slices and ensuring style consistency across long-range slices. To overcome these limitations, we propose the Style-aware Bidirectional Stream Diffusion Model, which ensures both inter-slice continuity and global style consistency with low computational cost. Specifically, we first utilize a Vector Quantized Variational Autoencoder to encode bidirectional adjacent CT slices into latent codes. A Neighbor Attention module is then introduced to capture transition patterns among these latent codes, ensuring structural continuity. To enhance style consistency, we further design a Prototype Prompting mechanism to construct a feature pool from long-range slices. Global style prototypes are extracted from the feature pool and dynamically integrated into the generation process, guiding the model's attention toward consistent stylistic features across slices. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods.

1 Introduction

Computed Tomography (CT) provides high-resolution anatomical details, while Positron Emission Tomography (PET) reveals metabolic activity. The fusion of CT and PET is crucial for early cancer screening and precise localization of malignant tumors along with their functional states [Mu *et al.*, 2020; Sujit *et al.*, 2024]. However, the high cost and additional radiation exposure associated with PET scanning limit its widespread application. In this context, image-to-image synthesis [Li *et al.*, 2025; Li *et al.*, 2024; Salehjahromi *et al.*, 2024] emerges as a promising alternative by leveraging conditional generative models to synthesize PET images directly from CT images (i.e., CT-to-PET synthesis).

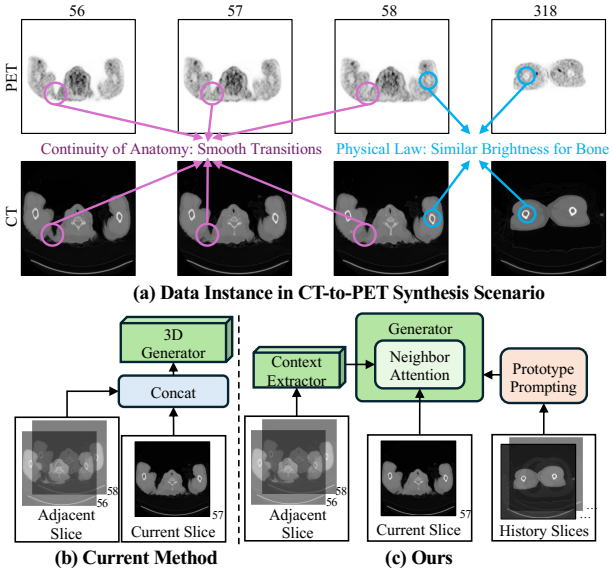


Figure 1: (a) CT and PET sequence data, where the numbers represent the index of the corresponding slice in the sequence. The purple circles highlight the smooth transitions between adjacent CT and PET slices, and the blue circles emphasize the similar brightness of bone regions across non-adjacent slices. (b) Illustration of the existing method. (c) The framework of the proposed method.

However, due to the sequential nature of CT and PET scans and the physical principles governing medical imaging, CT-to-PET synthesis faces two major challenges. First, continuous anatomical structures lead to strong structural correlation between adjacent slices (analogous to the content coherence between consecutive video frames). This necessitates the model to ensure smooth transitions between generated adjacent PET slices, as illustrated by slices 56, 57, and 58 in Fig. 1(a). Second, the imaging physics of CT and PET result in consistent imaging styles (e.g., brightness) for the same tissue type across non-adjacent slices, analogous to the object consistency maintained when the same object appears in non-adjacent video frames. For instance, as illustrated by slices 58 and 318 in Fig. 1(a), bone tissue appears bright in CT scans due to its high density, and similarly appears bright

in PET scans due to its low metabolic activity. Overall, these two challenges can be attributed to the broader issue of correlation: the first concerns the smooth transition arising from adjacent slice correlations, while the second involves the style consistency stemming from non-adjacent slice correlations.

Prior studies have explored advanced 2D diffusion models to enhance the fine-grained details of synthesized 2D slice while maintaining efficiency [Nguyen *et al.*, 2025]. However, in the absence of contextual information from adjacent slices, purely 2D models struggle to ensure inter-slice consistency. Conversely, as illustrated in Fig. 1(b), [Salehjahromi *et al.*, 2024] adopts 3D models [Wang *et al.*, 2024a; Wang *et al.*, 2024b] to process volumetric inputs and explicitly capture correlations across adjacent slices, but at the cost of substantially increased computational complexity. Fundamentally, existing methods face an inherent trade-off: 2D models offer computational efficiency but limited awareness of inter-slice correlations, whereas 3D models effectively model inter-slice correlations but incur prohibitive computational overhead.

To break the dilemma faced by existing methods while addressing both correlation challenges under acceptable computational cost, this work focuses on enhancing 2D slice generation by integrating information from both adjacent and non-adjacent slices. Specifically, this work propose a Style-aware Bidirectional Stream Diffusion Model (SBSDM), which integrates two key components, namely the Neighbor Attention module and the Prototype Prompting mechanism (see Fig. 1(c)). On one hand, to tackle the smooth transition challenge efficiently, we employ a lightweight Vector Quantized Variational Autoencoder (VQVAE) [Van Den Oord *et al.*, 2017] as a context extractor to capture stream transition information from bidirectional adjacent slices. The stream transition information is then integrated into the diffusion generation process for the current slice via the proposed Neighbor Attention module. Let D denote the computational cost of VQVAE feature extraction and C represent the cost of generating the current slice. The total computation of our model is $D \times (L-1) + C$, where L is the sequence length. When $D \ll C$ (facilitated by the lightweight VQVAE), we achieve substantial computational savings. On the other hand, to ensure style consistency across non-adjacent slices, we introduce a global Prototype Prompting mechanism. It aggregates features from all previously processed slices into a feature bank, from which global style prototypes are extracted and dynamically integrated into the generation process. This mechanism requires only minimal computational resources for storing and updating the feature bank, yet provides effective global style guidance throughout the synthesis process.

Our main contributions are summarized as follows:

- To ensure smooth transition, we develop a lightweight context extractor with a Neighbor Attention, enabling efficient extraction and integration of inter-slice transition patterns.
- To preserve style consistency, we propose a Prototype Prompting mechanism that stores global stylistic information and dynamically incorporates it into the generation process.
- Extensive experiments demonstrate that our method outperforms existing 3D methods while maintaining computational costs comparable to 2D models.
- We validate our SBSDM on the lung nodule benign–malignant classification task, demonstrating that preserving slice-sequence consistency effectively improves classification performance.

2 Related Works

Image-to-Image Synthesis in Medical Imaging: In medical imaging scenarios, image-to-image synthesis has shown promising results. For instance, [McNaughton *et al.*, 2023] employed a UNet architecture [Ronneberger *et al.*, 2015] to achieve bidirectional mapping between MRI and CT; Zhou *et al.* [Zhou *et al.*, 2020] and Yang *et al.* [Yang *et al.*, 2020] further validated the capability of deep neural networks in image synthesis across different MRI sequences. Chen *et al.* [Chen *et al.*, 2018] utilized Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2020] to achieve super-resolution from low-resolution MRI to high-resolution MRI. Additionally, Tan *et al.* [Tan *et al.*, 2024] explored synthetic methods for generating X-ray images from CT scans. Beyond the focus on morphological imaging modalities, the recent work [Nguyen *et al.*, 2025] has attempted to use diffusion models for CT-to-PET synthesis. However, it overlooked the sequential nature of PET imaging, resulting in failures to maintain cross-slice coherence and style consistency. On the other hand, [Salehjahromi *et al.*, 2024] proposed concatenating adjacent slices as input to help the model perceive transition information between adjacent slices. Yet, constructing 3D data through concatenation and using 3D models [Wang *et al.*, 2024a; Liu *et al.*, 2021] incurs significant computational costs [Zhou *et al.*, 2023], and the processable sequence length is constrained by GPU memory limitations, making it difficult for the model to maintain global style consistency.

Sequential Correlation Modeling: To capture sequential correlations in 3D data with either temporal or spatial dimensions, common approaches employ 3D CNN [Tran *et al.*, 2015] or 3D Transformers [Mao *et al.*, 2021] for automatic modeling of sequential correlations. For example, [Zhao *et al.*, 2024] utilized 3D CNN for the fusion and classification of 3D medical images, while [Wang *et al.*, 2024b] explored the combination of diffusion models with 3D models for CT volume generation. However, the practical application of 3D models is severely constrained by their high computational cost. To address this, prior work [Bertasius *et al.*, 2021] proposed divided spatial and temporal attention to model inter-frame correlations using 2D representations, thus avoiding the use of 3D models. [Liang *et al.*, 2024] introduced a feature bank and fusion mechanism to cache unidirectional temporal information, enabling real-time video generation based on 2D diffusion models [Song *et al.*, 2021]. In addition to architectural innovations, [Sun *et al.*, 2023] applied data augmentation techniques to compress 3D data into 2.5D representations for CT synthesis. Nevertheless, these prior efforts focus on limited sequence lengths or unidirectional temporal correlations, overlooking the need for modeling correlations

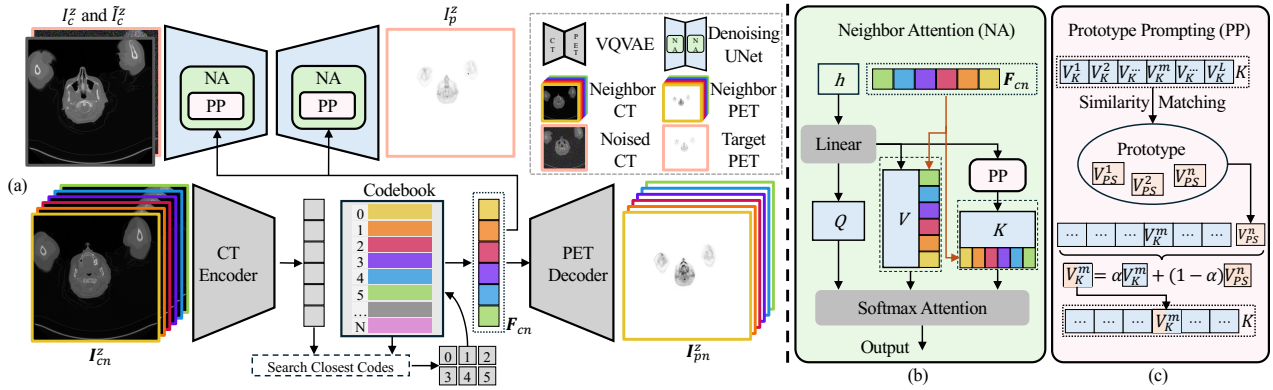


Figure 2: Illustration of the proposed model. The left subfigure (a) illustrates the overall computational workflow of the proposed model, while subfigures (b) and (c) depict the Neighbor Attention module and Prototype Prompting mechanism, respectively.

between bidirectional adjacent slices in medical image generation scenarios, and lacking the capability to capture long-range correlations between non-adjacent slices.

3 Methodology

3.1 Overview

In this work, CT-to-PET synthesis is formulated as a sequential generation task. Specifically, the model input consists of two components: 1) the current CT slice denoted as I_c^z, where the superscript z indicates the positional index within the CT scan sequence; and 2) the bidirectional adjacent slices represented by $\boldsymbol{I}_{cn}^z = \{I_c^{z-w}, I_c^{z-w+1}, \cdots, I_c^{z+w}\} \setminus \{I_c^z\}$, where $2w$ determines the number of adjacent slices considered. Correspondingly, the model architecture follows a dual-stream design. The first stream, illustrated in the lower part of Fig. 2(a), employs a lightweight context extractor based on VQVAE. The VQVAE is pretrained on paired CT and PET slices, and its encoder is then used to extract neighbor latent codes $\boldsymbol{F}_{cn}$ from $\boldsymbol{I}_{cn}^z$. As shown in the upper part of Fig. 2(a), the second stream is a diffusion-based synthesis pathway that receives latent codes $\boldsymbol{F}_{cn}$, the current CT slice I_c^z and its noisy version $\tilde{I}_c^z$ as inputs, and produces the synthetic PET slice I_p^z. Additionally, the core modules of the diffusion model, namely the Neighbor Attention (NA) module and the Prototype Prompting (PP) mechanism, are presented in Fig. 2(b) and (c), respectively. The NA module extracts transition patterns from $\boldsymbol{F}_{cn}$ to ensure smooth transition, while the PP mechanism adaptively integrates cross-slice style information into the key K of the NA module, thereby enhancing global style consistency.

3.2 Style-aware Bidirectional Stream Diffusion Model

The UNet [Ronneberger *et al.*, 2015; Rombach *et al.*, 2022] architecture with vanilla attention modules [Vaswani *et al.*, 2017] is employed as the backbone of our model. The core design of our model consists of two key components: 1) A NA module, which captures transition patterns from neighboring latent codes provided by a lightweight context extractor; 2) A PP mechanism, which extracts style prototypes from a feature pool and dynamically integrates them into the generation process.

Neighbor Attention for Smooth Transition

A lightweight context extractor is designed to capture neighbor features from bidirectional adjacent slices to facilitate smooth transitions in generated slices. Specifically, a quantized autoencoder is used as the context extractor [Van Den Oord *et al.*, 2017]. Following VQVAE [Van Den Oord *et al.*, 2017], the input I_c is first encoded by encoder E_{VQ} into features: $F_h = E_{VQ}(I_c) \in \mathbb{R}^d$. Then, the closest c_k to F_h is retrieved from the codebook $\mathcal{C} = \{c_k \in \mathbb{R}^d\}_{k=1}^{N_c}$ (N_c is the size of the codebook) as the quantized feature $F_c \in \mathbb{R}^d$:

$$F_c = \arg\min_{c_k} \|F_h - c_k\|_2 . \tag{1}$$

After obtaining F_c, the decoder D_{VQ} is used to reconstructs target domain images I_p^{vq} given F_c: $I_p^{vq} = D_{VQ}(F_c)$. We first train the VQVAE model on the training set $\mathcal{D}_{tr}$. For training details, please refer to [Van Den Oord *et al.*, 2017].

Subsequently, as shown in Fig. 2, the trained encoder (i.e., E_{VQ}) of the VQVAE is utilized to extract features from adjacent slices. For the CT slice I_c^z at index z, its adjacent sequence can be represented as:

$$\boldsymbol{I}_{cn}^z = \{I_c^{z-w}, I_c^{z-w+1}, \cdots, I_c^{z+w}\} \setminus \{I_c^z\}, \tag{2}$$

and the corresponding adjacent PET sequence is denoted as $\boldsymbol{I}_{pn}^z$. The latent codes of this adjacent sequence are denoted as:

$$\boldsymbol{F}_{cn} = \{F_c^{z-w}, F_c^{z-w+1}, \cdots, F_c^{z+w}\} \setminus \{F_c^z\}, \tag{3}$$

where $F_c^{z-w} = \arg\min_{c_k} \|E_{VQ}(I_c^{z-w}) - c_k\|_2$. The computational cost in obtaining $\boldsymbol{F}_{cn}$ primarily consists of two components: the forward computation cost of the VQVAE encoder E_{VQ} and the lookup computation cost from the codebook $\mathcal{C}$. By designing a lightweight encoder architecture, we can effectively extract contextual information from bidirectional adjacent slices with relatively low computational overhead. The dynamic balance between model performance and computational complexity will be thoroughly analyzed in the experimental section.

Although the VQVAE is capable of generating PET images from CT images, its architecture inherently limits the generation of high-quality data [Xiao *et al.*, 2022]. Therefore, the VQVAE is utilized solely to obtain the neighbor latent codes. Next, we aim to leverage the attention mechanism to automatically extract contextual information (i.e., transition patterns) from neighbor latent codes $\boldsymbol{F}_{cn}$ that are beneficial to the synthesis of the current slice. Specifically, we develop a brand new Neighbor Attention (NA) module to replace all the vanilla attention module in the backbone UNet.

As shown in Fig. 2, the noised input CT image is denoted as $\tilde{I}_c^z$ (obtained by the forward diffusion process), which is fed into the diffusion model. Then, the input representation to the NA module is denoted as $h \in \mathbb{R}^{L\times d}$. In addition, the neighbor latent codes is also fed into the NA module. In this module, a linear layer is utilized to map h into the query, key and value (i.e., $Q, K, V \in \mathbb{R}^{L\times d}$). The output of the NA module is calculated by:

$$\text{softmax}(\frac{Q\cdot[K,\boldsymbol{F}_{cn}]^\top}{\sqrt{d}})[V,\boldsymbol{F}_{cn}], \tag{4}$$

where $\boldsymbol{F}_{cn} \in \mathbb{R}^{2w\times d}$ is the neighbor latent code, which conveys bidirectional stream information from adjacent slices into the generation process of the current slice. $[\cdot,\cdot]$ represents the concatenation operation. Q, K, and V are representations of the current slice, while $\boldsymbol{F}_{cn}$ is representations of the bidirectional adjacent slices. The matrix multiplication between Q and $[K, \boldsymbol{F}_{cn}]$ produces a cross-slice similarity matrix, which is then passed through a softmax function to obtain attention weights. These attention weights act as dynamic feature selection gates, guiding the extraction of useful information from $[V, \boldsymbol{F}_{cn}]$ for synthesizing the current slice. It should be noted that the PP mechanism processing key representations in Fig. 2(b) is omitted here, as it will be elaborated in subsequent sections.

Prototype Prompting for Style Consistency

The proposed NA module focuses on extracting transition patterns from adjacent slices, but it fails to capture globally common visual styles due to the limited sequence length of adjacent slices. Therefore, a prototype prompting module is designed to tackle the challenge of style consistency. The solution lies in enabling the model to perceive consistent patterns across long-sequence slice representations. Inspired by the memory bank used in continual learning tasks [Liang and Li, 2024; Wang *et al.*, 2022], we develop a dynamic feature bank to store features of previously processed slices by the model.

Since the proposed diffusion model incorporates a core module: Neighbor Attention, the feature bank is specifically designed to handle the representations involved in this Neighbor Attention (NA) module. Taking the key representations K in one NA module as an example, we first initialize a zero matrix as the feature bank $P_K = \mathbf{0}_{d\times d}$ and set its initial size to zero $L_{PK} = 0$. Subsequently, by leveraging the property that slices within a batch are randomly sampled during training, we perform an averaging operation along the batch dimension to obtain a cross-slice representation: $K_{avg} = \frac{1}{N_b}\sum_{i=1}^{N_b} K_i \in \mathbb{R}^{L\times d}$, where K_i is the key representation (i.e., K mentioned above) of a single slice, and N_b is the batch size. L is the length of the representation, and d denotes the feature dimension. Next, K_{avg} is used to update the feature bank:

$$\begin{aligned} P_K &= \frac{L_{PK}P_K + K_{avg}^\top K_{avg}}{L_{PK}+L}, \\ L_{PK} &= L_{PK} + L, \end{aligned} \tag{5}$$

where $K_{avg}^\top K_{avg} \in \mathbb{R}^{d\times d}$ is the cross slice feature space. As the forward pass is performed on each data batch, the computed K_{avg} continuously updates P_K, gradually accumulating cross-slice features.

The feature bank is designed to cache cross-slice features. However, the feature bank stores a large number of cross-slice features. To further extract common style information while eliminating redundant features, we perform cluster analysis on the feature bank and use the cluster centroids as style prototypes. Specifically, we introduce a parameter-free clustering method: Gaussian Kernel-based Mean Shift [Cheng, 1995], which effectively distills prototypes from P_K: $P_S = \text{MS}(P_K) \in \mathbb{R}^{N_P\times d}$, where N_P is the number of prototypes (i.e., cluster centers). $\text{MS}(\cdot)$ represent the function of the Mean Shift.

Prototypes P_S are then merged with the key representations, thereby infusing the key representations with style information. Specifically, we design a Prototype Prompting mechanism. The feature K is represented as vector form $[V_K^1, V_K^2, \cdots, V_K^L]$, where each vector $V_K^m \in \mathbb{R}^d$, and $m \in [1, 2, \cdots, L]$. Similarly, $P_S = [V_{PS}^1, V_{PS}^2, \cdots, V_{PS}^{N_p}]$, where $V_{PS}^n \in \mathbb{R}^d$, and $n \in [1, 2, \cdots, N_p]$. Then, the cosine similarity between each V_K and V_{PS} is calculated as: $\frac{V_K\cdot V_{PS}}{\|V_K\|_2\|V_{PS}\|_2}$. For a feature vector V_K^m, we denote the feature vector in P_S that has the largest similarity with V_K^m as V_{PS}^n. If the similarity between V_K^m and V_{PS}^n exceeds a threshold θ, V_{PS}^n is fused into V_K^m:

$$V_K^m = \alpha V_K^m + (1-\alpha)V_{PS}^n, \tag{6}$$

where α is used to stabilize numerical computations. When V_K exhibits high similarity with V_{PS}, fusing V_{PS} can introduce style information while maintaining the stability of V_K. Conversely, when the similarity between V_K and V_{PS} is low, V_K primarily reflects the specific characteristics of the current slice. Forcibly fusing V_{PS} in such cases may lead to feature distortion. Therefore, the fusion operation is performed only when the similarity is sufficiently high.

It should be noted that the feature bank and Prototype Prompting mechanism are illustrated using the key representation in the NA module as an example, but they are also applicable to the value representation V and the NA outputs. Overall, the proposed diffusion model consists of two core components: the Neighbor Attention module, which integrates neighboring latent codes to ensure smooth transitions, and the Prototype Prompting module, which incorporates global style information to maintain style consistency in CT-to-PET synthesis.

Although several individual components in our framework are inspired by prior studies, the core contribution of this

Methods	Model Type	FDG-PET-CT-Lesions					In-house				
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIPs ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIPs ↑
CycleGAN	GAN	27.8741	0.9336	0.0967	121.5031	0.9259	26.7951	0.9126	0.1142	158.7106	0.9039
RegGAN		28.2336	0.9354	0.0867	110.1570	0.9362	27.1357	0.9204	0.1042	134.3641	0.9135
3D-RegGAN		28.3125	0.9364	0.0871	112.2016	0.9447	27.6561	0.9242	0.0959	124.8776	0.9282
DDBM-UNet	Diffusion	30.0545	0.9442	0.0528	60.9886	0.9611	29.2218	0.9385	0.0670	74.8731	0.9481
ResShift-UNet		30.0735	0.9457	0.0535	61.1946	0.9623	29.3577	0.9405	0.0673	75.7172	0.9492
DDBM-Freq		30.1320	0.9449	0.0517	59.6257	0.9646	29.4361	0.9403	0.0624	72.6610	0.9516
ResShift-Freq		30.1531	0.9465	0.0529	59.6926	0.9659	29.5273	0.9421	0.0629	72.7268	0.9528
DDBM-3DDiff		30.8574	0.9517	0.0385	56.1839	0.9727	30.1693	0.9486	0.0416	60.5662	0.9684
ResShift-3DDiff		30.8632	0.9521	0.0381	55.5741	0.9733	30.1841	0.9497	0.0417	60.5769	0.9688
DDBM-3DUNet		30.8847	0.9524	0.0377	55.0834	0.9759	30.0238	0.9472	0.0435	62.1771	0.9665
ResShift-3DUNet		30.9135	0.9539	0.0387	55.2120	0.9757	30.0662	0.9469	0.0433	62.3361	0.9671
VQVAE	Autoencoder	28.6825	0.9413	0.0761	91.3548	0.9341	28.0432	0.9314	0.0908	111.0392	0.9337
Ours	Ours	**31.9824**	**0.9627**	**0.0349**	**51.5145**	**0.9846**	**31.1292**	**0.9579**	**0.0367**	**53.5682**	**0.9740**

Table 1: Performance comparison on the task of synthetic PET from CT. The optimal performance and suboptimal performance are highlighted using **bold** and blue, respectively.

work lies not in isolated module invention, but in revealing and formulating the unique correlation mechanisms underlying CT-to-PET synthesis and designing a unified diffusion framework specifically tailored to these medical imaging characteristics.

Diffusion Training Paradigm

To train the proposed diffusion model, we adopt the same training objective as in [Yue *et al.*, 2023]:

$$\mathcal{L}_{diff} = \sum_{t} \|\mathcal{U}(\tilde{I}_c^z, I_c^z, \boldsymbol{F}_{cn}, t) - I_p^z\|_2^2, \tag{7}$$

where $\tilde{I}_c^z = (1-\eta_t)I_p^z + \eta_t I_c^z + \kappa\sqrt{\eta_t}\epsilon$ is the noised sample, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. t is the time step. η_t and κ are hyperparameters proposed in [Yue *et al.*, 2023]. $\mathcal{U}(\cdot)$ denotes our proposed diffusion model.

4 Experiments

4.1 Experimental Setup

Datasets: A large-scale publicly available dataset, FDG-PET-CT-Lesions [Gatidis *et al.*, 2022], is utilized to evaluate the effectiveness of the proposed model. This dataset comprises 1014 consecutive whole-body PET/CT scans. The dataset is split into 709 for training, 102 for validation, and 203 for testing, resulting in 248,657, 33,920, and 72,570 paired CT-PET slices for each subset, respectively. To further assess the model's generalizability, we additionally employ an in-house dataset consisting of 215 PET/CT scans. The in-house dataset is partitioned into 150 for training, 22 for validation, and 43 for testing, yielding 45,906, 6,930, and 13,639 paired CT-PET slices, respectively.

Evaluation Setting: We adopt four metrics for assessing individual slices quality: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Fréchet Inception Distance (FID), and the perceptual metric (LPIPS). Additionally, average cosine similarity (denoted as CLIPs) between CLIP embeddings of consecutive slices [Ceylan *et al.*, 2023] is used to evaluate the smooth transition (i.e., temporal coherency). Among these metrics, higher values of PSNR, SSIM, and CLIPs indicate better generation quality, while lower values of FID and LPIPS signify superior results.

Implementation Details: The VQVAE architecture follows the design in [Zhou *et al.*, 2022]. The codebook size is set to 1024. For the diffusion model, the shifting sequence η_t and the hyper-parameter $\kappa = 2$ follow the setting in [Yue *et al.*, 2023]. The number of diffusion time steps is set to 4 to ensure efficient and rapid image generation. The design of the UNet references Stable Diffusion [Rombach *et al.*, 2022]. The dimensionality of the feature vectors d is set to 640. The AdamW optimizer is used to train both the VQVAE and the UNet, with the batch size and learning rate set to 96 and 5e-5, respectively, for both models. The hyperparameters w, α, and θ are determined via parameter search and set to 10, 0.5, and 0.9, repsectively.

Baselines: To evaluate the effectiveness of our proposed method, we select the following baseline models: 1) CycleGAN [Zhu *et al.*, 2017]: A direct implementation of the pioneering method [Salehjahromi *et al.*, 2024]; 2) RegGAN [Kong *et al.*, 2021]: A GAN model improved for medical imaging characteristics; 3) 3D-RegGAN: Replacing the 2D generator of RegGAN with a 3D UNet [Çiçek *et al.*, 2016] architecture; 4) DDBM-UNet: Modifies the training paradigm of ResShift-UNet to DDBM, representing a straightforward implementation of the current state-of-the-art method [Nguyen *et al.*, 2025]; 5) ResShift-UNet [Yue *et al.*, 2023]: Combining the advanced diffusion training paradigm ResShift with the standard diffusion model UNet [Rombach *et al.*, 2022]; 6) DDBM-Freq and 7) ResShift-Freq: Introducing high-frequency components improves the quality of generated details [Li *et al.*, 2024]; 8) DDBM-3DDiff and 9) ResShift-3DDiff: A 3D CNN is employed to transform the high-frequency components [Shang *et al.*, 2024]; 10) DDBM-3DUNet and 11) ResShift-3DUNet: The UNet in 4) and 5) is replaced with the 3D UNet; 11) VQVAE: an autoencoder-based image synthesis method [Van Den Oord *et al.*, 2017].

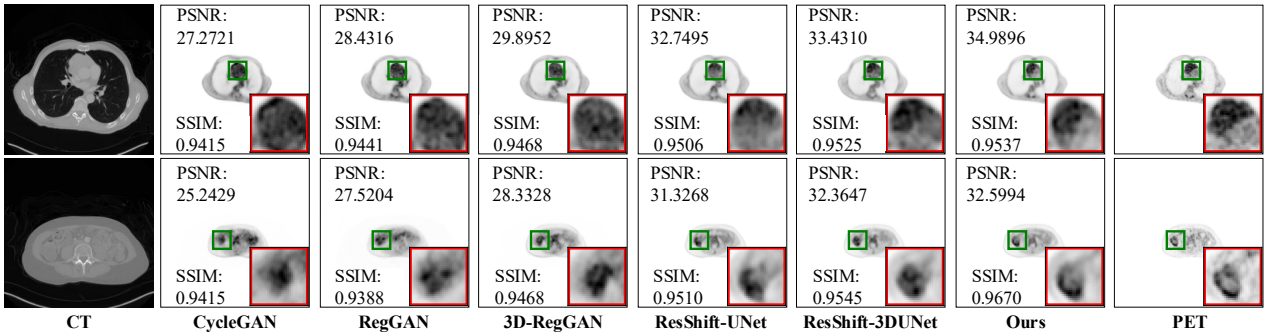


Figure 3: Visualization of generated results comparing different methods. The first and last columns represent the input CT and the ground truth PET, respectively. The content inside the red boxes represents a magnified view of the region enclosed by the green box.

4.2 Comparison with the State-of-the-Art Methods

Table 1 summarizes the quantitative performance of our method and baseline models on both datasets under identical training settings. The results support the following key observations: 1) VQVAE alone fails to produce satisfactory results, highlighting its limited generative capability. This justifies our design choice of using VQVAE solely for extracting neighbor latent codes. 2) Our method surpasses even the strongest 3D diffusion-based baselines. Fig. 3 presents a qualitative comparison. It can be observed that our 2D UNet-based framework achieves superior performance even compared to ResShift-3DUNet.

4.3 Ablation Study

NA	PP	FDG-PET-CT-Lesions				
		PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	CLIPs ↑
✗	✗	30.0735	0.9457	0.0535	61.1946	0.9623
✗	✓	30.6579	0.9511	0.0471	57.0172	0.9652
✓	✗	30.9320	0.9542	0.0392	54.4575	0.9780
✓	✓	31.9824	0.9627	0.0349	51.5145	0.9846

Table 2: Ablation analysis of the proposed components to assess their contributions to performance improvement. "NA" represents the Neighbor Attention module. "PP" denotes the Prototype Prompting. The ablated and non-ablated modules are represented by ✗ and ✓, respectively.

Comprehensive ablation studies are conducted to validate the effectiveness of the proposed NA module and PP mechanism. The results of the quantitative analysis are presented in Table 2, where we assess the contribution of NA and PP by selectively removing each component from the full model. The results show that removing either the NA or PP module leads to a notable performance degradation, clearly demonstrating the importance of both components.

To intuitively assess the effectiveness of the NA module in addressing the smooth transition challenge, Fig. 4 presents generation results for a continuous sequence with and without the NA module. These results demonstrate that the NA module enhances the model's ability to perceive anatomical transitions between adjacent slices.

Figure 4: Qualitative ablation analysis of the proposed NA module. The PET images in adjacent columns correspond to neighboring slices in the original data. The first row displays the sequence generation results after removing the NA module, the second row shows the generation results of the complete model, and the third row represents the real PET sequence.

Furthermore, we conduct a case study to evaluate the effectiveness of the PP module in maintaining style consistency. As shown in Fig. 5, removing the PP module (second column, W/O PP) causes low-metabolism visceral regions to be incorrectly synthesized as high-metabolism areas. In contrast, the full model (third column, Ours) better preserves the PET color style of low-metabolism regions, demonstrating the PP module's effectiveness in maintaining global style consistency across synthesized slices.

4.4 Computational Efficiency Analysis

To evaluate the effect of VQVAE parameter scale, we vary the number of residual blocks in the encoder and measure the corresponding performance. As shown in Fig. 6(a), performance becomes stable when the number of residual blocks exceeds 12. Therefore, we set the number of residual blocks to 12, resulting in an encoder with approximately 81M parameters and 71G FLOPs, achieving a balance between efficiency and performance. In addition, Fig. 6(b) shows that our model requires significantly lower FLOPs than 3DUNet as

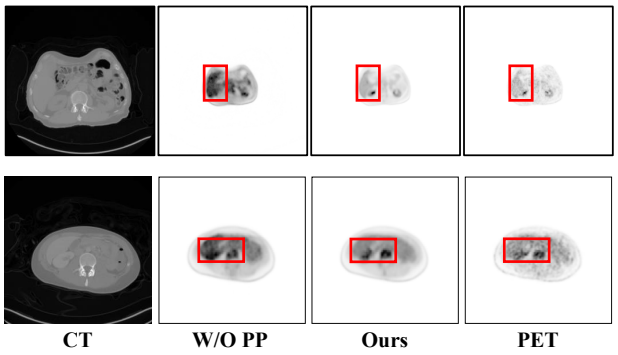
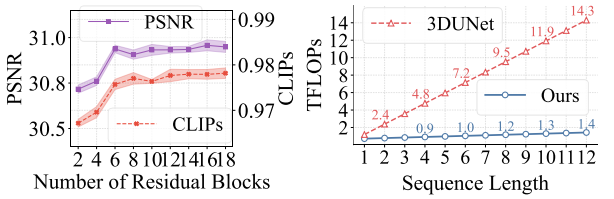


Figure 5: Qualitative ablation analysis on the PP method. The red boxes highlight regions with significant differences.

the sequence length increases. Furthermore, the inference efficiency comparison in Fig. 7, conducted on an NVIDIA RTX 5880 Ada Generation GPU, demonstrates that our method achieves inference speed comparable to 2D models such as ResShift-UNet while providing superior performance.



(a) Performance Scaling with VQVAE (b) Computational Cost Comparison

Figure 6: (a) Model performance w.r.t. the number of residual blocks in E_{VQ}. (b) FLOPs comparison between our model and 3DUNet across varying sequence lengths.

4.5 Clinical Utility Analysis of Synthetic PET

We evaluate the clinical utility of the synthesized PET images by applying them to a lung nodule malignancy classification task. Specifically, we first use the model trained on the FDG-PET-CT-Lesions dataset to synthesize PET slices from CT slices in the LIDC-IDRI dataset [Armato III *et al.*, 2011] (79 and 78 cases are used for training and validation, respectively). A binary classification model [Sibille *et al.*, 2020] is then trained under three different settings: 1): using only CT; 2): using CT combined with PET synthesized by ResShift-3DUNet; 3), 4), and 5): using CT combined with PET synthesized by three variants of our method: without NA, without PP, and the full model, respectively. The results in Table 3 verify the effectiveness and clinical application potential of each module we proposed. Due to limited pathologically confirmed cases in LIDC-IDRI, our validation is only a preliminary test of synthetic PET's potential in downstream tasks.

We employ two radiologists from commercial medical institutions to evaluate the quality of the generated PET slices. Specifically, we randomly select 50 PET slices from the test set and mix them with corresponding PET slices generated by different methods. Without being informed about the existence of synthetic PET images, the radiologists are asked

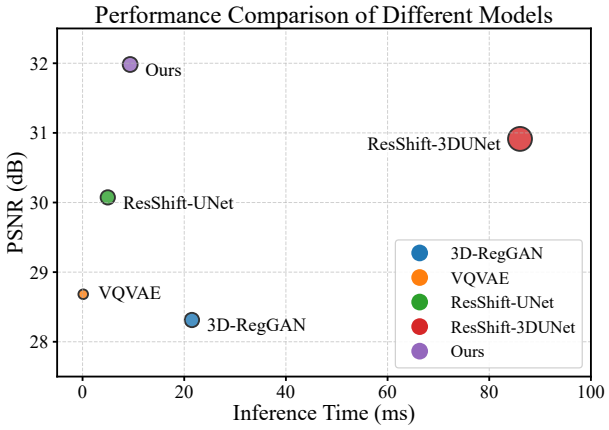


Figure 7: Efficiency analysis of existing methods vs. our proposed method. The horizontal axis represents inference time (in milliseconds), while the vertical axis denotes PSNR. The area of each circle corresponds to the model's parameter count.

to score all PET scans on a scale from 1 to 5, with higher scores indicating better image quality. The average score is then computed and summarized in Table 4. It can be observed that the PET generated by our method achieved the closest scores to the real PET.

Metrics	w/o PET	3DUNet	Ours (✗ NA)	Ours (✗ PP)	Ours
Acc	0.6026	0.6410	0.6282	0.6538	0.6667
Recall	0.5949	0.6364	0.6293	0.6515	0.6586

Table 3: Experimental results of the nodule classification task.

Method	3D-RegGAN	ResShift-Unet	ResShift-3DUNet	Ours	Groud-Truth
Score	2.32	2.94	3.14	3.45	4.04

Table 4: Comparative quality assessment of different generation methods based on radiologists' evaluation.

5 Conclusion

In this work, we propose a Style-aware Bidirectional Stream Diffusion Model to address the CT-to-PET synthesis task. Specifically, we introduce the NA module to capture transition patterns between adjacent slices and the PP mechanism to ensure global style consistency across synthesized PET images. Extensive experiments on both public and private datasets demonstrate the effectiveness of our approach. The CLIPs metric quantitatively verifies the NA module's ability to ensure smooth inter-slice transitions, while a case study confirms the PP mechanism's contribution to preserving global style consistency. In future work, we plan to explore the physiological foundations underlying deep learning-based CT-to-PET synthesis.

References

[Armato III *et al.*, 2011] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. *Med. Phys.*, 38(2):915–931, Feb. 2011.

[Bertasius *et al.*, 2021] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. Int. Conf. Mach. Learn.*, pages 1–13, 2021.

[Ceylan *et al.*, 2023] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 23206–23217, 2023.

[Chen *et al.*, 2018] Yuhua Chen, Feng Shi, Anthony G. Christodoulou, Yibin Xie, Zhengwei Zhou, and Debiao Li. Efficient and accurate mri super-resolution using a generative adversarial network and 3d multi-level densely connected network. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pages 91–99, 2018.

[Cheng, 1995] Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 17(8):790–799, Aug. 1995.

[Çiçek *et al.*, 2016] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pages 424–432, 2016.

[Gatidis *et al.*, 2022] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Sci. Data*, 9(1):601, Oct. 2022.

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

[Kong *et al.*, 2021] Lingke Kong, Chenyu Lian, Detian Huang, ZhenJiang Li, Yanle Hu, and Qichao Zhou. Breaking the dilemma of medical image-to-image translation. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 34, pages 1964–1978, 2021.

[Li *et al.*, 2024] Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Trans. Med. Imaging*, 43(3):980–993, Mar. 2024.

[Li *et al.*, 2025] Yisheng Li, Yishan Wang, Baiying Lei, and Shuqiang Wang. Scdm: Unified representation learning for eeg-to-fnirs cross-modal generation in mi-bcis. *IEEE Trans. Med. Imaging*, 44(6):2384–2394, Jun. 2025.

[Liang and Li, 2024] Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 23638–23647, 2024.

[Liang *et al.*, 2024] Feng Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, and Diana Marculescu. Looking backward: Streaming video-to-video translation with feature banks. In *Proc. Int. Conf. Learn. Represent.*, pages 1–14, 2024.

[Liu *et al.*, 2021] Yucheng Liu, Yulin Liu, Rami Vanguri, Daniel Litwiller, Michael Liu, Hao-Yun Hsu, Richard Ha, Hiram Shaish, and Sachin Jambawalikar. 3d isotropic super-resolution prostate mri using generative adversarial networks and unpaired multiplane slices. *J. Digit. Imaging*, 34:1199–1208, Sep. 2021.

[Mao *et al.*, 2021] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3164–3173, 2021.

[McNaughton *et al.*, 2023] Jake McNaughton, Samantha Holdsworth, Benjamin Chong, Justin Fernandez, Vickie Shim, and Alan Wang. Synthetic mri generation from ct scans for stroke patients. *BioMedInformatics*, 3(3):791–816, Sep. 2023.

[Mu *et al.*, 2020] Wei Mu, Lei Jiang, JianYuan Zhang, Yu Shi, Jhanelle E Gray, Ilke Tunali, Chao Gao, Yingying Sun, Jie Tian, Xinming Zhao, et al. Non-invasive decision support for nsclc treatment using pet/ct radiomics. *Nat. Commun.*, 11(1):5228, Oct. 2020.

[Nguyen *et al.*, 2025] Dac Thai Nguyen, Trung Thanh Nguyen, Huu Tien Nguyen, Thanh Trung Nguyen, Huy Hieu Pham, Thanh Hung Nguyen, Thao Nguyen Truong, and Phi Le Nguyen. Ct to pet translation: A large-scale dataset and domain-knowledge-guided diffusion approach. In *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, pages 1498–1507, 2025.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10684–10695, 2022.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, pages 234–241, 2015.

[Salehjahromi *et al.*, 2024] Morteza Salehjahromi, Tatiana V Karpinets, Sheeba J Sujit, Mohamed Qayati, Pingjun Chen, Muhammad Aminu, Maliazurina B Saad, Rukhmini Bandyopadhyay, Lingzhi Hong, Ajay Sheshadri, et al. Synthetic pet from ct improves diagnosis and prognosis for lung cancer: Proof of concept. *Cell Rep. Med.*, 5(3):101463, Mar. 2024.

[Shang *et al.*, 2024] Shuyao Shang, Zhengyang Shan, Guangxing Liu, LunQian Wang, XingHua Wang, Zekai

Zhang, and Jinglin Zhang. Resdiff: Combining cnn and diffusion model for image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8975–8983, 2024.

[Sibille *et al.*, 2020] Ludovic Sibille, Robert Seifert, Nemanja Avramovic, Thomas Vehren, Bruce Spottiswoode, Sven Zuehlsdorff, and Michael Schäfers. 18f-fdg pet/ct uptake classification in lymphoma and lung cancer by using deep convolutional neural networks. *Radiology*, 294(2):445–452, Dec. 2020.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proc. Int. Conf. Learn. Represent.*, pages 1–12, 2021.

[Sujit *et al.*, 2024] Sheeba J Sujit, Muhammad Aminu, Tatiana V Karpinets, Pingjun Chen, Maliazurina B Saad, Morteza Salehjahromi, John D Boom, Mohamed Qayati, James M George, Haley Allen, et al. Enhancing nsclc recurrence prediction with pet/ct habitat imaging, ctdna, and integrative radiogenomics-blood insights. *Nat. Commun.*, 15(1):3152, Apr. 2024.

[Sun *et al.*, 2023] Bin Sun, Shuangfu Jia, Xiling Jiang, and Fucang Jia. Double u-net cyclegan for 3d mr to ct image synthesis. *Int. J. Comput. Assist. Radiol. Surg.*, 18(1):149–156, Aug. 2023.

[Tan *et al.*, 2024] Lixing Tan, Shuang Song, Kangneng Zhou, Chengbo Duan, Lanying Wang, Huayang Ren, Linlin Liu, Wei Zhang, and Ruoxiu Xiao. Multi-view x-ray image synthesis with multiple domain disentanglement from ct scans. In *Proc. ACM Int. Conf. Multimed.*, pages 3209–3218, 2024.

[Tran *et al.*, 2015] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 4489–4497, 2015.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 1–10, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 6000–6010, 2017.

[Wang *et al.*, 2022] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 631–648, 2022.

[Wang *et al.*, 2024a] Yan Wang, Yanmei Luo, Chen Zu, Bo Zhan, Zhengyang Jiao, Xi Wu, Jiliu Zhou, Dinggang Shen, and Luping Zhou. 3d multi-modality transformer-gan for high-quality pet reconstruction. *Med. Image Anal.*, 91:102983, Jan. 2024.

[Wang *et al.*, 2024b] Yuran Wang, Zhijing Wan, Yansheng Qiu, and Zheng Wang. Devil is in details: Locality-aware 3d abdominal ct volume generation for self-supervised organ segmentation. In *Proc. ACM Int. Conf. Multimed.*, pages 10640–10648, 2024.

[Xiao *et al.*, 2022] Zhisheng Xiao, Karsten Kreis, and Arash Vahdat. Tackling the generative learning trilemma with denoising diffusion GANs. In *Proc. Int. Conf. Learn. Represent.*, pages 1–15, 2022.

[Yang *et al.*, 2020] Qianye Yang, Nannan Li, Zixu Zhao, Xingyu Fan, Eric I-Chao Chang, and Yan Xu. Mri cross-modality image-to-image translation. *Sci. Rep.*, 10(1):3753, Feb. 2020.

[Yue *et al.*, 2023] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Resshift: Efficient diffusion model for image super-resolution by residual shifting. In *Proc. Adv. Neural Inf. Process. Syst.*, volume 36, pages 13294–13307, 2023.

[Zhao *et al.*, 2024] Shanbo Zhao, Meihui Zhang, Xiaoqin Zhu, Junjie Li, Yunyun Duan, Zhizheng Zhuo, Yaou Liu, and Chuyang Ye. Medmm: A multimodal fusion framework for 3d medical image classification with multigranular text guidance. In *Proc. Int. Conf. Big Data Comput. Commun.*, pages 42–49, 2024.

[Zhou *et al.*, 2020] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: Hybrid-fusion network for multi-modal mr image synthesis. *IEEE Trans. Med. Imaging*, 39(9):2772–2781, Sep. 2020.

[Zhou *et al.*, 2022] Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 30599–30611, 2022.

[Zhou *et al.*, 2023] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Xiaoguang Han, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Volumetric medical image segmentation via a 3d transformer. *IEEE Trans. Image Process.*, 32:4036–4045, Jul. 2023.

[Zhu *et al.*, 2017] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2223–2232, 2017.