A Prior-information-guided Residual Diffusion Model for Multi-modal PET Synthesis from MRI

Zaixin Ou[1], **Caiwen Jiang**[1], **Yongsheng Pan**[2], **Yuanwang Zhang**[1], **Zhiming Cui**[1] and **Dinggang Shen**[1,3,4*]

[1]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University, Shanghai, China
[2]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China
[3]Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[4]Shanghai Clinical Research and Trial Center, Shanghai, China
dgshen@shanghaitech.edu.cn

Abstract

Alzheimer's disease (AD) leads to abnormalities in various biomarkers (i.e., amyloid-β and tau proteins), which makes PET imaging (which can detect these biomarkers) essential in AD diagnosis. However, the high radiation risk of PET imaging limits its scanning number within a short period, presenting challenges to the joint multi-biomarker diagnosis of AD. In this paper, we propose a novel unified model to simultaneously synthesize multi-modal PET images from MRI, to achieve low-cost and time-efficient joint multi-biomarker diagnosis of AD. Specifically, we incorporate residual learning into the diffusion model to emphasize inter-domain differences between PET and MRI, thereby forcing each modality to maximally reconstruct its modality-specific details. Furthermore, we leverage prior information, such as age and gender, to guide the diffusion model in synthesizing PET images with semantic consistency, enhancing their diagnostic value. Additionally, we develop an intra-domain difference loss to ensure that the intra-domain differences among synthesized PET images closely match those among real PET images, promoting more accurate synthesis, especially at the modality-specific information. Extensive experiments conducted on the ADNI dataset demonstrate that our method achieves superior performance both quantitatively and qualitatively compared to the state-of-the-art methods. All codes for this study have been uploaded to GitHub (https://github.com/Ouzaixin/ResDM).

1 Introduction

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disorder, and early diagnosis is crucial for effective disease management and improving the patient's life quality [Provost *et al.*, 2021]. Given that AD causes changes in multiple biomarkers, the National Institute on Aging and Alzheimer's Association proposed the joint multi-biomarker (amyloid-β, tau, and neurodegeneration) diagnostic approach, called ATN analysis [Jack Jr *et al.*, 2018]. Conducting ATN analysis requires collecting the patient's Aβ-PET, tau-PET, and MRI to acquire information on amyloid-β (Aβ) deposition, tau proteins, and neurodegeneration, respectively. However, the high costs (especially radiation risks) of PET imaging limit the feasibility of conducting multiple PET scans (i.e., Aβ-PET and tau-PET) within a short period, posing challenges to real-time ATN analysis and thus restricting diagnostic accuracy. To achieve cost-effective and real-time ATN analysis, a feasible approach is to simultaneously synthesize corresponding Aβ-PET and tau-PET images from Magnetic Resonance Imaging (MRI).

Synthesizing multi-modal PET images from MRI is not a straightforward task, as it is a one-to-many generative problem, especially when distinguishing similar modalities such as Aβ-PET and tau-PET. The most direct approach for such tasks is to use multiple models for different generative tasks [Lan *et al.*, 2021; Deng *et al.*, 2022]. While simple, this method overlooks the interdependencies inherent in the related generative tasks and incurs computational redundancy due to the need to train multiple models. Consequently, researchers are increasingly leaning towards using a unified model to handle multiple generative tasks simultaneously. One approach in unified model design is to have multiple generative tasks share parts of the model architecture. For instance, [Jiang *et al.*, 2023] employs a shared encoder and two task-specific decoders to simultaneously synthesize two types of dual-energy CT images from single-energy CT images. This method effectively reduces the training burden and captures the dependencies between different generative tasks to some extent. To further enhance the performance of unified models, some studies have explored leveraging text control, enabling different generative tasks through the same model pathway. This strategy significantly reduces model parameters and adequately captures the dependencies between tasks.

Diffusion models are a recently popular type of text-controlled unified model. They consist of a process that gradually adds noise to an input image until it becomes a

*Corresponding author

pure noise, and also a denoising process that continuously removes noise from the noised samples to recover the input image. The diffusion-based models decompose the generative task into a series of simple steps, allowing for precise control of the generation process using text information. However, they are often guided by simple textual forms, such as "Aβ-PET" and "tau-PET", neglecting some highly relevant prior information associated with the images. Besides, our task involves generating highly similar images (i.e., Aβ-PET and tau-PET) within the same domain from MRI. Traditional diffusion models struggle to differentiate between these two PET modalities effectively, ultimately limiting their performance of generation. Furthermore, despite using the same network to synthesize PET images of different modalities, they do not impose further constraints on the synthesized images of different PET modalities, i.e., their relationships and discrepancy, resulting in suboptimal performance.

To tackle these issues, we propose a novel unified model to simultaneously synthesize multi-modal PET images from MRI. The core of our method is incorporating residual learning into the diffusion model to emphasize inter-domain differences between PET and MRI, rather than their nearly identical anatomical structures. The residual learning not only helps the model differentiate two similar PET modalities but also facilitates a meticulous exploration of distinctive features between the PET domain and the MRI domain. To improve the diagnostic value of synthesized PET images, we take the subject's age and gender as the prior information to guide the diffusion model in synthesizing PET images with semantic consistency. Besides, to constrain the PET images from different modalities, we develop an intra-domain difference loss to ensure that the intra-domain differences among synthesized PET images closely match those among real PET images. The proposed novel components and loss functions can effectively synthesize multi-modal PET images and boost the usability of our model in real-world clinical scenarios.

Our main contributions are summarized as follows:

- Incorporating residual learning into the diffusion model to emphasize inter-domain differences between PET and MRI, thereby forcing each modality to maximally reconstruct its modality-specific details for AD diagnosis.
- Leveraging prior information, such as age and gender, to guide the diffusion model in synthesizing PET images with semantic consistency, enhancing their diagnostic value.
- Developing an intra-domain difference loss to ensure that the intra-domain differences among synthesized PET images closely match those among real PET images, promoting more accurate synthesis.

2 Related Works

2.1 Multi-Model PET Synthesis from MRI

The high costs of acquiring PET images have limited their clinical use, leading to research efforts aimed at reducing these costs through image generation technologies. Among these efforts, generating corresponding PET images from MRI has garnered significant attention. This research can be divided into three categories based on the modality of the target PET images: 1) Synthesis of FDG-PET from MRI [Li *et al.*, 2014; Wang *et al.*, 2018; Lan *et al.*, 2021; Pan *et al.*, 2021; Hu *et al.*, 2021]; 2) Synthesis of Aβ-PET from MRI [Kang *et al.*, 2020; Kimura *et al.*, 2020; Jin *et al.*, 2023; Vega *et al.*, 2023]; 3) Synthesis of tau-PET from MRI [Sun *et al.*, 2022; Lee *et al.*, 2022; Jang *et al.*, 2023]. Current studies mainly focused on synthesizing a single modality of PET images, yet accurate AD diagnosis (ATN analysis) requires multi-modal PET images. Consequently, our objective is to develop a method for synthesizing multi-modal PET images from MRI, offering a rapid and cost-effective solution for precise multi-biomarker joint AD diagnosis.

2.2 Residual Learning in Image Generation

Residual learning, initially developed to solve gradient explosion in image recognition tasks [He *et al.*, 2016], focuses on enabling models to learn the differences (i.e., residuals) between inputs and outputs, rather than directly learning the mapping relationship. Due to its effectiveness in making model learning more efficient and targeted, researchers try to apply residual learning to generative tasks as well. For example, [Jifara *et al.*, 2019] used it in autoencoders for the image denoising task, allowing the network to more effectively restore image details. [Gao *et al.*, 2020] proposed a two-level residual learning CNN for image super-resolution tasks to achieve high-resolution images with high-frequency components. [Duan *et al.*, 2021] introduced a residual Generative Adversarial Network (GAN) for the cross-modal translation task to help the model more efficiently learn complex image distributions. [Huang *et al.*, 2023] incorporated the residual learning into the Vision transformer (ViT) to ease the training of deeper ViT and significantly alleviate the degradation problem. Inspired by these works, we incorporate the concept of residual learning into the diffusion process of diffusion models, by proposing a residual diffusion model that can better differentiate similar data distributions.

2.3 Text-Guided Synthesis

Text-guided synthesis is a flexible and effective method for controlling image generation, aimed at producing images that are highly consistent with the associated textual descriptions. Early methods [Choi *et al.*, 2018; Tang *et al.*, 2018; Duan *et al.*, 2021] simply transform input texts into discrete labels, which are then used to control image generation. Although this approach could control image generation to some extent, much of the semantic information from the original text was significantly lost in the process of converting to discrete labels. To address this, researchers have tried to use convolutional network-based feature extraction modules [Rombach *et al.*, 2022] to extract semantic features from texts. However, due to the limited scale of the feature extraction modules, these methods could only handle simple texts and lack flexibility. Fortunately, the development of large-scale pre-trained language models [Radford *et al.*, 2021; Nichol *et al.*, 2021; Ramesh *et al.*, 2022] has recently made it possible to extract semantic information from input texts more effectively and flexibly. In this paper, we design a prior-information-guided module based on the Contrastive

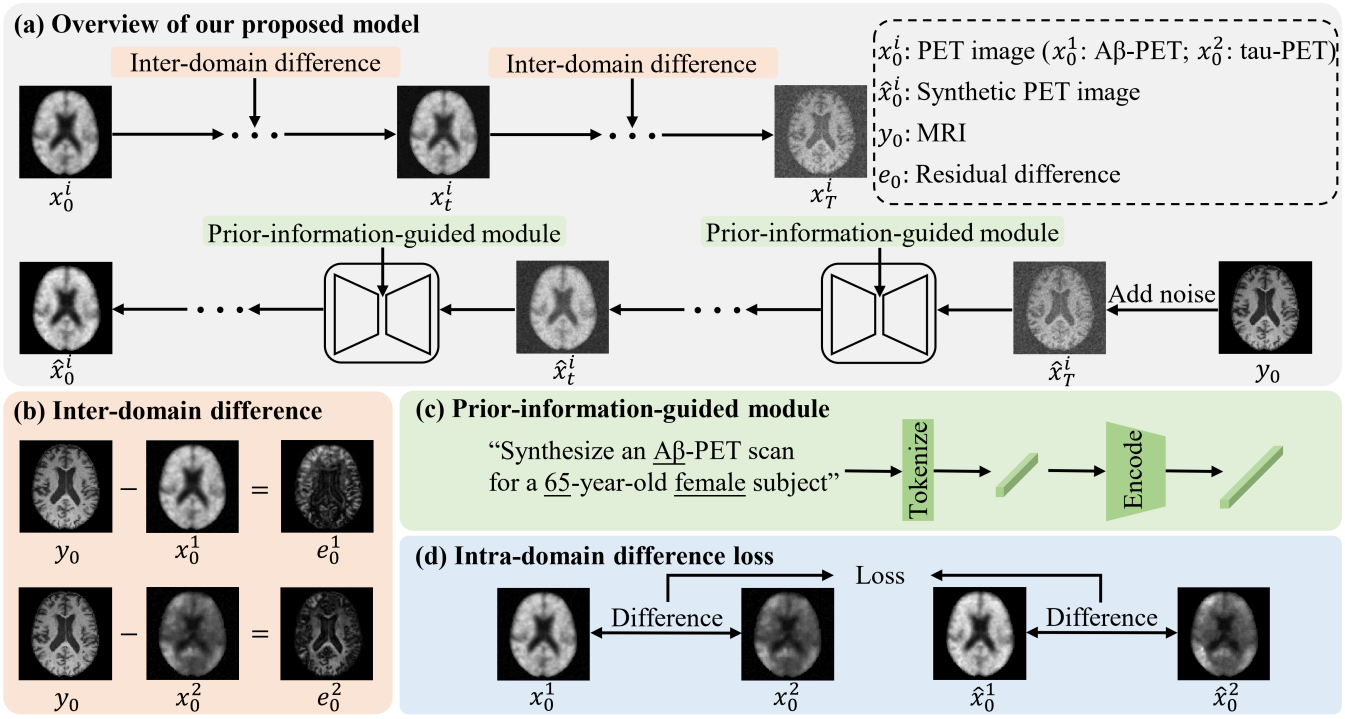

Figure 1: Overview of our proposed method. (a) The pipeline of residual diffusion model, (b) The inter-domain differences, (c) Details of the prior-information-guided module, and (d) The intra-domain difference loss.

Language-Image Pre-training model (CLIP [Radford *et al.*, 2021]), which efficiently and flexibly handles various information such as image categories and individual attributes. This module is then used to guide the image generation process.

3 Methodology

The overview of our proposed method is shown in Figure 1 (a). Given an input PET image x_0^i (where $i = 1$ for Aβ-PET, $i = 2$ for tau-PET), it first undergoes a series of transformations into noisy MRI x_T^i through the residual diffusion process, gradually merging inter-domain differences. The noisy MRI x_T^i then undergoes a series of reconstruction steps under the control of the prior-information-guided module until it is restored to the input PET image. We also design an intra-domain difference loss to constrain the generation between different PET modalities to improve generation performance.

3.1 Residual Diffusion Model

MRI and corresponding multi-modal PET images (i.e., Aβ and tau-PET) have a substantial proportion of similarities, as they are acquired from the same individual. Using traditional diffusion models to learn the mapping from MRI to different PET modalities can lead to significant redundancy, and the model might struggle to learn the critical parts (i.e., the differences between MRI and different PET modalities, as shown in Figure 1 (b)). Therefore, inspired by the concept of residual learning, we design a residual diffusion strategy to enable the model to directly learn the differences between MRI and different PET modalities.

In the forward process, we transform the input PET image x_0 into a noisy MRI x_T through a series of transformations. To achieve this, we first extract the inter-domain difference $e_0 = y_0 - x_0$ between MRI y_0 and the input PET image x_0, and then apply a shifting sequence $\{\eta_t\}_{t=1}^T$ to add this difference e_0 to the input PET image. Thus, the intermediate version x_t can be modeled as:

$$q(x_t|x_0, y_0) = \mathcal{N}(x_t; x_0 + \eta_t e_0, \Sigma) \tag{1}$$

where $\eta_1 \to 0$, and $\eta_T \to 1$.

In the reverse process, we progressively restore the noisy MRI x_T (i.e., the noisy version of MRI y_0) back into the input PET image x_0 through a series of reconstruction steps. Representing the posterior distribution as $p(x_0|y_0)$, this process can be formalized as follows:

$$p(x_0|y_0) = \int p(x_T|y_0) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, y_0) \mathrm{d}x_{1:T} \tag{2}$$

where we assume the marginal distributions of x_T converges to $\mathcal{N}(x_T; y_0, \Sigma)$, i.e., $p(x_T|y_0) \approx \mathcal{N}(x_T; y_0, \Sigma)$. Following commonly used strategies in diffusion models [Dhariwal and Nichol, 2021; Song *et al.*, 2020], we adopt the assumption of $p_\theta(x_{t-1}|x_t, y_0) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, y_0, t), \Sigma_\theta(x_t, y_0, t))$. The $\Sigma_\theta(x_t, y_0, t)$ can be viewed as a fixed variance and $\mu_\theta(x_t, y_0, t)$ can be reparameterized as follows:

$$\mu_{\theta(x_t, y_0, t)} = \frac{\eta_{t-1}}{\eta_t} x_t + \frac{\eta_t - \eta_{t-1}}{\eta_t} f_\theta(x_t, y_0, t) \tag{3}$$

where f_θ is a deep neural network with learnable parameter θ, aiming to predict x_0. Consequently, the reconstruction loss

of our model is defined as follows:

$$\mathcal{L}_{rec} = \min_{\theta} \sum_{t} \|f_\theta(x_t, y_0, t) - x_0\|_2^2 . \quad (4)$$

3.2 Prior-Information-Guided Module

In this work, we aim to accomplish the generation of both Aβ-PET and tau-PET images by using a single model. To achieve this, we guide each generation with corresponding textual information to produce the expected PET modality. Meanwhile, considering that our generated images are ultimately for diagnostic tasks, we also wish to integrate clinically relevant information, such as age and gender, into the generation process to enhance the diagnostic value of the images. To this end, we utilize a large-scale pre-trained language model, namely CLIP, to design a prior-information-guided module. This module can effectively and flexibly process these textual information to guide image generation.

The details of the prior-information-guided module are provided in Figure 1 (c). Specifically, this module takes a piece of text information as input and then tokenizes the text into concept tokens with a length of 77. The concept tokens are then processed through a text encoder based on ViT-B/32 CLIP to extract a textual vector. The extracted textual vector is subsequently used to guide a series of image reconstruction processes. Noting that considering the original pre-trained CLIP model is designed for natural images and corresponding texts, we employ fine-tuning techniques on the final transformer layer of CLIP to ensure that the model's understanding of the input clinical information can align with the target image generation task.

During each reconstruction step, we guide the reconstruction by fusing the textual vector and the image feature maps extracted from the bottleneck of each reconstruction network, serving as the input for the next layer. We utilize cross-attention to fuse the textual vector and image feature maps, which can be mathematically represented as $Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V$ with

$$Q = W_Q^{(i)} \cdot z_{image}, K = W_K^{(i)} \cdot z_{text}, V = W_V^{(i)} \cdot z_{text} \quad (5)$$

where z_{image} is the flattened image feature maps from the bottleneck layer, and z_{text} represents the textual vector.

3.3 Intra-Domain Difference Loss

Since Aβ-PET and tau-PET are highly similar, it's quite challenging for the model to accurately reconstruct different modalities of PET images from noisy MRI using only the image category information provided by the prior-information-guided module. Thus, we further design an intra-domain difference loss to constrain the synthesis of PET images across different modalities.

Specifically, in each reconstruction step, we employ the intra-domain difference loss to ensure that the differences between predicted Aβ-PET and tau-PET images at multiple levels align with the differences between the real Aβ-PET and tau-PET images, as shown in Figure 1 (d). This can be formulated as follows:

$$\mathcal{L}_{high} = \min_{\theta} \sum_{t} \left\| |\hat{x}_0^1 - \hat{x}_0^2| - |x_0^1 - x_0^2| \right\|_2^2 \quad (6)$$

	AD, N = 210	MCI, N = 656	CN, N = 608
Age	74.58±8.09	71.66±7.48	71.37±6.63
Sex (Female/Male)	85/125	297/359	357/251
Years of education	15.66±2.61	16.19±2.55	16.68±2.37
MMSE	23.01±2.35	27.96±1.81	29.09±1.13
Aβ SUVR	1.40±0.22	1.22±0.23	1.12±0.18
Tau SUVR	1.60±0.37	1.33±0.30	1.18±0.12

Table 1: Demographics for the collected dataset, with some data being shown as mean±std.

where $\hat{x}_0^1$ and $\hat{x}_0^2$ are the synthesized Aβ-PET and tau-PET images at each time step t, and x_0^1 and x_0^2 denote real Aβ-PET and tau-PET images, respectively.

By combining the reconstruction loss ($\mathcal{L}_{rec}$) and the intra-domain difference loss ($\mathcal{L}_{high}$), the comprehensive objective function is formulated as:

$$\mathcal{L} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{high} \quad (7)$$

where λ is a hyper-parameter to balance different loss terms.

4 Experiments

4.1 Dataset and Metric

To validate our proposed method, we collect 1274 sets of Aβ-PET, tau-PET, and T1-weighted MRI from the public Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, with 1020 sets used for training and 254 sets for testing. Each set of images is acquired from the same individual, along with corresponding AD diagnostic labels, and aligned to a common space by affine registration. To eliminate anisotropy, we resample all images to a size of $128 \times 128 \times 128$ with a voxel spacing of 1.5 mm^3. Additional details about the collected dataset can be found in Table 1.

We comprehensively evaluated the experimental results from both the reconstruction and diagnosis aspects. In the reconstruction evaluation, we evaluated the quality of the synthesized images using Mean Absolute Error (MAE), Structural Similarity Index (SSIM), and Peak Signal-to-Noise Ratio (PSNR). As for the diagnosis evaluation, we assess the AD diagnostic value of the synthesized images through the AD diagnosis task using the pre-trained disease diagnosis model in [Pan *et al.*, 2021], and quantify the diagnostic outcomes using Accuracy (ACC), F1 Score (F1S), and the Area Under the Receiver Operating Characteristic Curve (AUC).

4.2 Comparison with State-of-the-Art Methods

We compare our proposed method with five state-of-the-art generative approaches, including: 1) Pix2Pix [Isola *et al.*, 2017], 2) SC-GAN [Lan *et al.*, 2021], 3) TransUNet [Chen *et al.*, 2021], 4) ResViT [Dalmaz *et al.*, 2022], and 5) DDPM [Ho *et al.*, 2020]. In our experiments, Pix2Pix, SC-GAN, and TransUNet use two structurally identical models to separately synthesize Aβ-PET and tau-PET images, whereas ResViT and DDPM employ a single model with control modules to generate different modality PET images. Specifically, ResViT utilizes an availability condition module to randomly mask the target modality for learning, enabling the

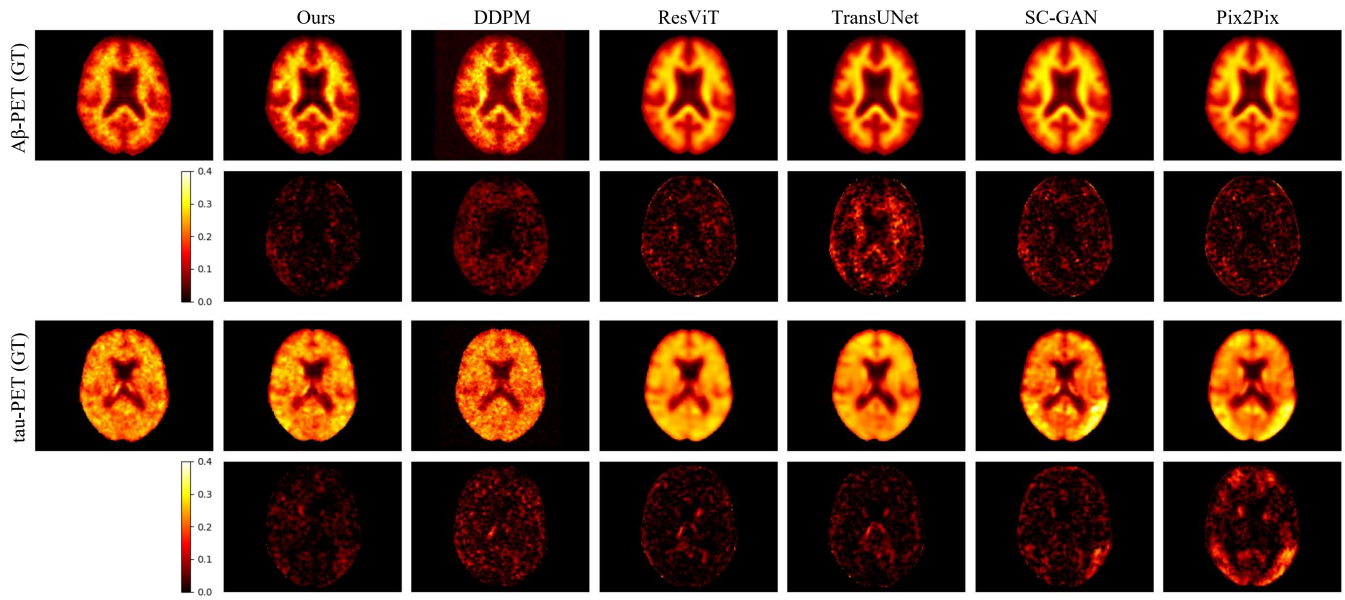

Figure 2: Top two rows: Ground-truth images and synthesized Aβ-PET images with error maps by six models on a typical subject from the test set. Bottom two rows: Ground-truth images and synthesized tau-PET images with error maps by the same six models.

	Synthetic Aβ-PET			Synthetic tau-PET		
	MAE(%)↓	PSNR(dB)↑	SSIM(%)↑	MAE(%)↓	PSNR(dB)↑	SSIM(%)↑
Pix2Pix	3.11±1.15	23.50±2.52	87.70±2.57	3.02±1.55	24.49±3.59	86.02±3.18
SC-GAN	2.75±0.95	24.37±2.38	89.76±2.36	2.76±1.74	25.49±3.95	88.15±3.62
DDPM	2.79±1.08	24.85±2.66	88.19±2.83	2.42±1.78	26.59±4.38	89.84±3.68
TransUNet	2.63±0.98	24.86±2.74	90.87±2.49	2.41±1.70	26.66±4.17	90.17±3.87
ResViT	2.54±0.92	25.20±2.51	91.05±2.28	2.37±1.89	26.91±4.46	90.19±4.19
Ours	**2.11±0.71**	**26.38±2.18**	**92.49±2.27**	**1.90±1.18**	**27.78±3.21**	**91.44±3.05**

Table 2: Quantitative evaluation of the synthetic images by six models. Results are listed as mean±std calculated across the test dataset.

synthesis of diverse target modalities during the inference stage. DDPM is modified by aligning MRI with noisy PET through channel concatenation and integrating text information (represented as discrete vectors) into the denoising process for multi-model PET synthesis, using a common approach as described in [Xie *et al.*, 2022; Lyu and Wang, 2022; Özbey *et al.*, 2023]. The quantitative and qualitative results are provided in Table 2 and Figure 2, respectively.

Quantitative Comparison: As shown in Table 2, our proposed method outperforms others, showcasing robust performance in generating both Aβ-PET and tau-PET images. Meanwhile, compared to the traditional diffusion model (i.e., DDPM), we achieve substantial improvements across all metrics. Specifically, for Aβ-PET generation, the improvements in MAE, PSNR, and SSIM are respectively 0.43%, 1.18dB, and 1.44%. For tau-PET generation, the improvements in MAE, PSNR, and SSIM are respectively 0.53%, 0.87dB, and 1.25%. This demonstrates the effectiveness of our targeted improvements to the design of the traditional diffusion model.

Qualitative Comparison: The visual results are provided in Figure 2. From the figure, it is evident that methods based on GAN, including ResViT, TransUNet, SC-GAN, and Pix2Pix, tend to synthesize overly smooth images. In contrast, the PET images generated by our proposed method exhibit the clearest critical details. Moreover, the PET images reconstructed by our approach show the smallest difference from the ground truth, as indicated by the lightest colors in the corresponding error maps. These findings collectively demonstrate that our method achieves superior performance to state-of-the-art methods.

4.3 Ablation Studies

To validate the different components of our proposed method, we use the residual diffusion model as a baseline and selectively incorporate intra-domain difference loss and a prior-information-guided module to form several variants. The results are provided in Table 3 and Figure 3. As indicated in Table 2, the baseline (1st row) shows improvements upon the addition of either intra-domain difference loss (2nd row) or prior-information-guided module (3rd row), and it achieves the best synthetic results when both are applied simultaneously (4th row). This validates the benefit of our designed intra-domain difference loss and prior-information-guided module in synthesizing PET images. Additionally,

	Synthetic Aβ-PET				Synthetic tau-PET			
	MAE(%)↓	PSNR(dB)↑	SSIM(%)↑	ACC(%)↑	MAE(%)↓	PSNR(dB)↑	SSIM(%)↑	ACC(%)↑
Base	2.52±0.92	25.27±2.54	91.41±2.10	80.31±2.19	2.38±1.73	26.82±4.32	90.30±3.85	81.84±4.80
w $\mathcal{L}_{high}$	2.14±0.98	25.78±2.51	92.07±2.47	80.66±1.54	2.12±1.82	27.49±3.93	90.92±3.71	82.04±4.74
w prior	2.17±0.86	25.98±2.70	92.18±2.32	81.74±1.59	2.00±1.83	27.57±3.73	91.10±3.56	82.68±5.08
Ours	**2.11±0.71**	**26.38±2.18**	**92.49±2.27**	81.90±1.74	**1.90±1.18**	**27.78±3.21**	**91.44±3.05**	83.06±6.16

Table 3: Quantitative evaluation of the synthetic images by our proposed model and its variants.

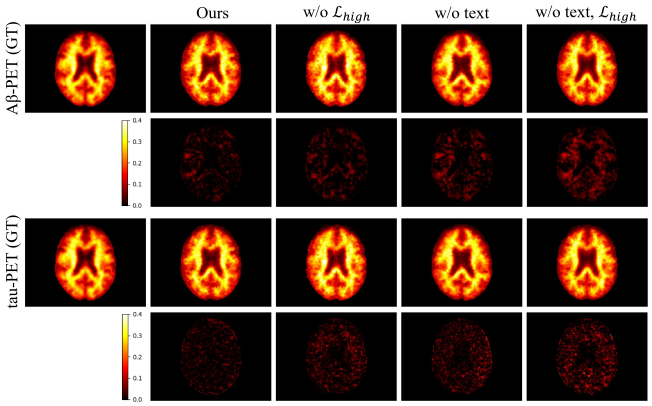

Figure 3: Top two rows: Ground-truth and synthesized Aβ-PET images with error maps by our proposed model and its variants on a typical subject from the test set. Bottom two rows: Ground-truth and synthesized tau-PET images with error maps by our proposed model and its variants.

the incorporation of the prior-information-guided module resulted in significant improvements in the positive and negative diagnoses of Aβ and tau proteins. This demonstrates that the utilized prior information enhances the synthesis of PET images with greater diagnostic value. Figure 3 illustrates that the synthesized PET images by integrating both intra-domain difference loss and prior-information-guided module matches best with the ground truth, further proving the effectiveness of each proposed component.

5 Discussions

5.1 The Role of Prior Information

To further explore the role of prior information in PET image generation, we design two experimental groups to evaluate the impact of two common types of prior information (namely, gender, and age) on PET image synthesis. For gender information, we create texts with correct gender, no gender, and incorrect gender, and then use these texts to guide the synthesis of PET images. The synthesized PET images are subjected to a gender classification task by using the classification model proposed in [Pan *et al.*, 2021]. The diagnostic results are provided in Figure 4. It can be observed that PET images synthesized with guidance from the correct gender information exhibit the strongest gender classification capability, followed by those synthesized with no gender information and incorrect gender information. This indirectly indicates the sensitivity of image synthesis to the guiding gender information, where correct gender guidance leads to images with greater diagnostic value.

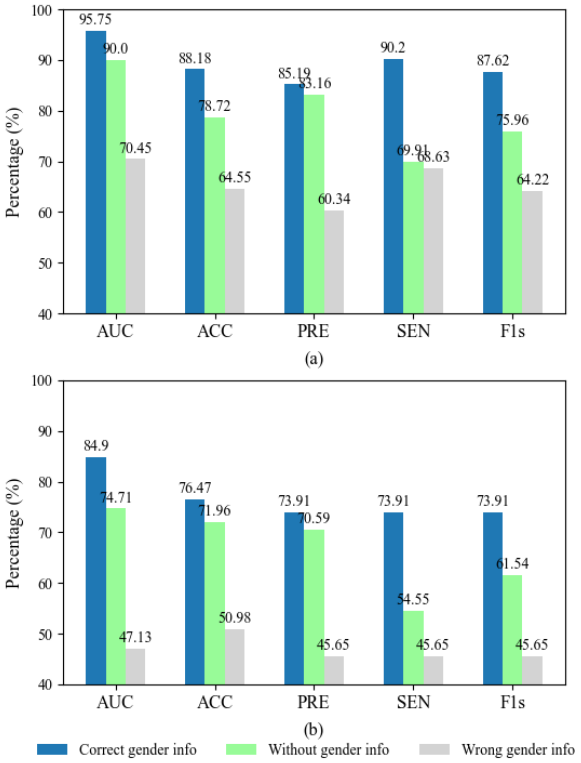

Figure 4: Conditioning analysis of gender information. (a) Performance metrics for synthetic Aβ-PET images under different conditions. (b) Performance metrics for synthetic tau-PET images under different conditions.

Similar to the assessment with gender information, we evaluate the sensitivity of synthesized images to age information under two scenarios, i.e., with and without age guidance. Specifically, we use a prediction model [Cole *et al.*, 2017] to estimate the actual age based on the generated PET images, with results presented in Figure 5. The results reveal that, under age-guided conditions, whether using generated Aβ-PET or tau-PET images for age prediction, the correlation between predicted and actual ages is significantly higher compared to the non-age-guided scenario. This result further demonstrates that our designed prior-information-guided module effectively enables the model to synthesize PET images highly semantically related to the corresponding textual information.

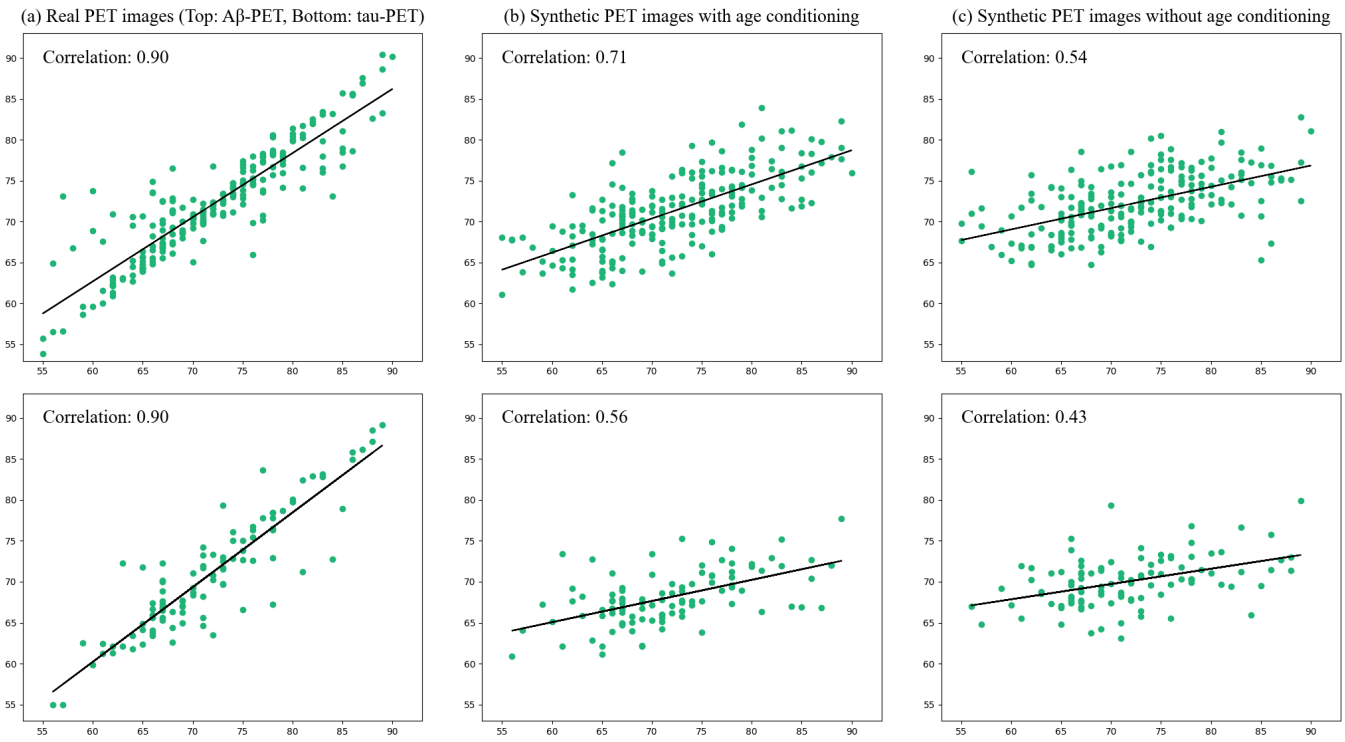

Figure 5: Conditioning analysis of age information. (a) Correlation between conditioning brain age and predicted brain age from real Aβ-PET (Top) and tau-PET (Bottom) images, respectively. (b) Correlation using Aβ-PET (Top) and tau-PET (Bottom) images with age conditioning. (c) Correlation using synthetic Aβ-PET (Top) and tau-PET (Bottom) images without age conditioning.

	Aβ status classification			tau status classification		
	AUC (%)↑	ACC (%)↑	F1S (%)↑	AUC (%)↑	ACC (%)↑	F1S (%)↑
SC-GAN	71.20±3.62	63.92±2.66	66.04±4.42	75.44±6.08	76.52±6.62	61.90±6.08
ResViT	83.12±1.97	76.82±1.79	75.34±4.11	82.92±3.10	77.80±5.26	63.28±3.76
TransUNet	85.76±1.24	77.52±0.70	77.34±2.43	87.08±2.48	79.46±5.77	66.98±8.35
Pix2Pix	87.96±1.96	78.86±2.61	79.40±3.72	85.66±4.99	81.46±5.15	69.24±6.05
DDPM	89.20±1.43	80.68±1.85	80.66±4.97	88.42±2.62	82.30±7.06	71.42±7.60
Ours	**90.74±1.50**	**81.90±1.74**	**82.74±3.68**	**90.02±1.88**	**83.02±6.16**	**76.67±6.00**
Real modality	98.96±0.22	94.18±0.77	94.12±1.19	96.90±0.22	92.32±4.20	86.44±7.83

Table 4: Quantitative evaluation of Aβ status and tau status classification. Results are listed as mean±std through five-fold cross-validation.

5.2 Diagnostic Value of Synthetic Images

Furthermore, we evaluate the diagnostic value of the generated images by diagnosing the status of Aβ and tau proteins using the generated Aβ-PET and tau-PET images. The diagnostic results using PET images generated by different methods are presented in Table 4. The statistical results show that diagnoses made using PET images generated by our method achieve results close to those using real images, with accuracies (ACC) of 81.90% and 83.02% for the positive and negative diagnoses of Aβ and tau proteins, respectively. Moreover, compared to other methods, our method achieves the best diagnostic results. These results suggest that our proposed method can synthesize PET images with exceptionally high diagnostic value, showing potential for application in clinical diagnosis.

6 Conclusions

In this paper, we propose a prior-information-guided residual diffusion model to simultaneously generate multi-modal PET images from MRI, addressing the challenge of acquiring multi-modal PET images within a short time. We specifically design three strategies tailored to this task, including 1) utilizing residual diffusion to enable the model to focus only on the inter-domain differences between MRI and PET, 2) embedding a prior-information-guided module for integrating image modality and subject attributes into the generation process, and 3) designing an intra-domain difference loss to constrain the differences among different PET modalities. Experiments on the ADNI dataset demonstrate that our method outperforms state-of-the-art approaches and shows potential for application in clinical auxiliary diagnosis.

Acknowledgments

This work was supported in part by National Natural Science Foundation of China (grant numbers 62131015, 62250710165), Science and Technology Commission of Shanghai Municipality (STCSM) (grant number 21010502600), STI2030-Major Project (No. 2022ZD0213100), and The Key R&D Program of Guangdong Province, China (grant number 2021B0101420006).

References

[Chen *et al.*, 2021] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[Choi *et al.*, 2018] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[Cole *et al.*, 2017] James H Cole, Rudra PK Poudel, Dimosthenis Tsagkrasoulis, Matthan WA Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[Dalmaz *et al.*, 2022] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. ResViT: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

[Deng *et al.*, 2022] Lizhen Deng, Chunming He, Guoxia Xu, Hu Zhu, and Hao Wang. Pcgan: A noise robust conditional generative adversarial network for one shot learning. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):25249–25258, 2022.

[Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[Duan *et al.*, 2021] Bin Duan, Wei Wang, Hao Tang, Hugo Latapie, and Yan Yan. Cascade attention guided residue learning GAN for cross-modal translation. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 1336–1343. IEEE, 2021.

[Gao *et al.*, 2020] Min Gao, Xian-Hua Han, Jing Li, Hui Ji, Huaxiang Zhang, and Jiande Sun. Image super-resolution based on two-level residual learning cnn. *Multimedia Tools and Applications*, 79:4831–4846, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Hu *et al.*, 2021] Shengye Hu, Baiying Lei, Shuqiang Wang, Yong Wang, Zhiguang Feng, and Yanyan Shen. Bidirectional mapping generative adversarial networks for brain MR to PET synthesis. *IEEE Transactions on Medical Imaging*, 41(1):145–157, 2021.

[Huang *et al.*, 2023] Guoxi Huang, Hongtao Fu, and Adrian G Bors. Masked image residual learning for scaling deeper vision transformers. *arXiv preprint arXiv:2309.14136*, 2023.

[Isola *et al.*, 2017] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[Jack Jr *et al.*, 2018] Clifford R Jack Jr, David A Bennett, Kaj Blennow, Maria C Carrillo, Billy Dunn, Samantha Budd Haeberlein, David M Holtzman, William Jagust, Frank Jessen, Jason Karlawish, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia*, 14(4):535–562, 2018.

[Jang *et al.*, 2023] Se-In Jang, Cristina Lois, Emma Thibault, J Alex Becker, Yafei Dong, Marc D Normandin, Julie C Price, Keith A Johnson, Georges El Fakhri, and Kuang Gong. TauPETGen: Text-conditional tau PET image synthesis based on latent diffusion models. *arXiv preprint arXiv:2306.11984*, 2023.

[Jiang *et al.*, 2023] Caiwen Jiang, Yongsheng Pan, Tianyu Wang, Qing Chen, Junwei Yang, Li Ding, Jiameng Liu, Zhongxiang Ding, and Dinggang Shen. S2DGAN: Generating dual-energy CT from single-energy CT for real-time determination of intracerebral hemorrhage. In *International Conference on Information Processing in Medical Imaging*, pages 375–387. Springer, 2023.

[Jifara *et al.*, 2019] Worku Jifara, Feng Jiang, Seungmin Rho, Maowei Cheng, and Shaohui Liu. Medical image denoising using convolutional neural network: a residual learning approach. *The Journal of Supercomputing*, 75:704–718, 2019.

[Jin *et al.*, 2023] Yan Jin, Jonathan DuBois, Chongyue Zhao, Liang Zhan, Audrey Gabelle, Neda Jahanshad, Paul M Thompson, Arie Gafson, and Shibeshih Belachew. Brain MRI to PET synthesis and amyloid estimation in Alzheimer's disease via 3D multimodal contrastive GAN. In *International Workshop on Machine Learning in Medical Imaging*, pages 94–103. Springer, 2023.

[Kang *et al.*, 2020] Hyeon Kang, Jang-Sik Park, Kook Cho, and Do-Young Kang. Visual and quantitative evaluation of amyloid brain PET image synthesis with generative adversarial network. *Applied Sciences*, 10(7):2628, 2020.

[Kimura *et al.*, 2020] Yuichi Kimura, Aya Watanabe, Takahiro Yamada, Shogo Watanabe, Takashi Nagaoka, Mitsutaka Nemoto, Koichi Miyazaki, Kohei Hanaoka, Hayato Kaida, and Kazunari Ishii. AI approach of cycle-consistent generative adversarial networks to synthesize

PET images to train computer-aided diagnosis algorithm for dementia. *Annals of Nuclear Medicine*, 34:512–515, 2020.

[Lan *et al.*, 2021] Haoyu Lan, Alzheimer Disease Neuroimaging Initiative, Arthur W Toga, and Farshid Sepehrband. Three-dimensional self-attention conditional GAN with spectral normalization for multimodal neuroimaging synthesis. *Magnetic resonance in medicine*, 86(3):1718–1733, 2021.

[Lee *et al.*, 2022] Jeyeon Lee, Brian J Burkett, Hoon-Ki Min, Matthew L Senjem, Ellen Dicks, Nick Corriveau-Lecavalier, Carly T Mester, Heather J Wiste, Emily S Lundt, Melissa E Murray, et al. Synthesizing images of tau pathology from cross-modal neuroimaging using deep learning. *bioRxiv*, pages 2022–09, 2022.

[Li *et al.*, 2014] Rongjian Li, Wenlu Zhang, Heung-Il Suk, Li Wang, Jiang Li, Dinggang Shen, and Shuiwang Ji. Deep learning based imaging data completion for improved brain disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part III 17*, pages 305–312. Springer, 2014.

[Lyu and Wang, 2022] Qing Lyu and Ge Wang. Conversion between CT and MRI images using diffusion and score-matching models. *arXiv preprint arXiv:2209.12104*, 2022.

[Nichol *et al.*, 2021] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[Özbey *et al.*, 2023] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

[Pan *et al.*, 2021] Yongsheng Pan, Mingxia Liu, Yong Xia, and Dinggang Shen. Disease-image-specific learning for diagnosis-oriented neuroimage synthesis with incomplete multi-modality data. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6839–6853, 2021.

[Provost *et al.*, 2021] Karine Provost, Leonardo Iaccarino, David N Soleimani-Meigooni, Suzanne Baker, Lauren Edwards, Udo Eichenlaub, Oskar Hansson, William Jagust, Mustafa Janabi, Renaud La Joie, et al. Comparing ATN-T designation by tau PET visual reads, tau PET quantification, and CSF PTau181 across three cohorts. *European journal of nuclear medicine and molecular imaging*, 48:2259–2271, 2021.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[Song *et al.*, 2020] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[Sun *et al.*, 2022] Yuqing Sun, Yong Liu, and Bing Liu. Predicting conversion to mild cognitive impairment in cognitively normal with incomplete multi-modal neuroimages. In *2022 10th International Conference on Bioinformatics and Computational Biology (ICBCB)*, pages 61–65. IEEE, 2022.

[Tang *et al.*, 2018] Hao Tang, Dan Xu, Wei Wang, Yan Yan, and Nicu Sebe. Dual generator generative adversarial networks for multi-domain image-to-image translation. In *Asian Conference on Computer Vision*, pages 3–21. Springer, 2018.

[Vega *et al.*, 2023] Fernando Vega, Abdoljalil Addeh, Aravind Ganesh, Eric E Smith, and M Ethan MacDonald. Image translation for estimating two-dimensional axial amyloid-beta PET from structural MRI. *Journal of Magnetic Resonance Imaging*, 2023.

[Wang *et al.*, 2018] Yan Wang, Luping Zhou, Biting Yu, Lei Wang, Chen Zu, David S Lalush, Weili Lin, Xi Wu, Jiliu Zhou, and Dinggang Shen. 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. *IEEE transactions on medical imaging*, 38(6):1328–1339, 2018.

[Xie *et al.*, 2022] Taofeng Xie, Chentao Cao, Zhuoxu Cui, Fanshi Li, Zidong Wei, Yanjie Zhu, Ye Li, Dong Liang, Qiyu Jin, Guoqing Chen, et al. Brain PET synthesis from MRI using joint probability distribution of diffusion model at ultrahigh fields. *arXiv preprint arXiv:2211.08901*, 2022.