STMDiff: Spatiotemporal Matching Diffusion Model for Dual-Time-Point Total-body PET/CT Imaging via Contrastive Learning

Wenbo Li[1], Zhenxing Huang[1], Lianghua Li[2], Chunyan Yang[1], Yihan Wang[3], Wenjian Qin[1], Na Zhang[1], Hairong Zheng[1], Dong Liang[1], Jianjun Liu[2], and Zhanli Hu[1,✉]

[1] Research Center for Medical AI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China.
zl.hu@siat.ac.cn

[2] Department of Nuclear Medicine, Ren Ji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai 201807, China.

[3] Central Research Institute, United Imaging Healthcare Group, Shanghai 201807, China.

Abstract. Total body PET/CT systems, which enable unprecedented image quality and ultrahigh sensitivity, are widely utilized for diagnosing and treating diseases like tumors. Unlike regular protocols, dual-time-point imaging (DTPI)– where patients undergo a dual PET/CT scan to enhance lesion contrast – exposes them to higher radiation doses due to an additional CT scan for PET attenuation correction and anatomical localization. To mitigate radiation exposure, we introduce STMDiff, a spatiotemporal matching diffusion model, which reuse CT images from first scanning time point for PET attenuation correction at second scanning time point. Spatiotemporal matching strategy implemented with contrastive learning aims to find the k-best-matched CT images, which enriches the multimodal features of STMdiff and bypasses the cross-modal registration, facilitating the generation of attenuation-corrected (AC) PET images alleviating alignment errors. Both qualitative and quantitative results illustrate that the AC PET images from STMDiff not only obtain the best quantitative scores (PSNR: 37.72 ± 6.85 dB; SSIM: 0.96 ± 0.03; RMSE: 2.35 ± 1.03), but also preserve metabolic information. Moreover, clinical assessment results show that the standardized uptake value (SUV) distribution of our method is more consistent with that of real AC PET images[4].

Keywords: Dual-Time-Point PET/CT Imaging · Spatiotemporal Matching · Contrastive Learning

1 Introduction

PET/CT systems, combining the functional and anatomical information, have been widely applied in oncology, cardiology, and neurology. Recently, the advent

[4] Our code is available at https://github.com/LEE12365/STMDiff

of total-body PET/CT systems have led to unprecedented levels of image quality and quantification accuracy with approximately 40-fold increased sensitivity [1, 2]. Conventional PET/CT data is usually collected at a specific time point, around 50 minutes after the tracer injection. Different from this regular scanning protocol, dual-time-point imaging (DTPI) involves two PET/CT scans at different time points (approximately 60 and 150 minutes), which result in the PET images obtained at second time point with reduced background activity, increasing the rate of lesions detection [3], as shown in Fig. 1. Although DTPI has many advantages in tumor diagnosis, staging, and therapeutic evaluation, it also limited to additional CT radiation caused by the second PET/CT scanning.

Attenuation correction (AC) has played an important role in PET/CT imaging. It could correct gamma-ray attenuation effects and improve the visual interpretation and quantification accuracy of PET images [4, 5]. For PET/CT systems, CT imaging not only provides high-resolution structural information, but also offer attenuation coefficients for PET attenuation correction. In this case, ionizing radiation from CT becomes a major source of these CT-based attenuation correction methods, especially in DTPI. Hence eliminating the need of a second CT scan in DTPI has become an urgent priority.

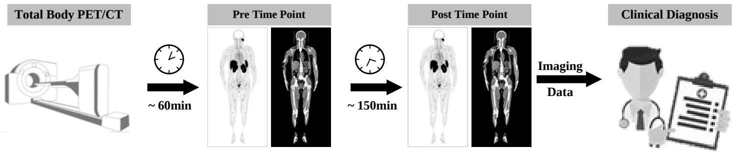

Fig. 1. The Illustration of DTPI Scanning Protocol.

Many deep learning (DL)-based studies commit to reduce CT radiation dose for PET attenuation correction[6, 7]. Some methods attempted to generate pseudo-CT (sCT) images from non-attenuation-corrected PET images (NAC PET). However, the presence of local or global estimation bias in sCT images may introduce quantitative errors. Some approaches tried to achieve CT-free attenuation correction, which directly synthesized AC PET images from NAC PET images [8–10]. Despite these methods could directly eliminate CT radiation, they are more sensitive to the image noise. Recently, diffusion models have greatly developed in high-quality image generation, which not only improve image generation quality but also enhance training stability [11–13]. Therefore, it is promising to apply the diffusion model to achieve the CT-free attenuation correction in DTPI for total-body PET/CT imaging.

With these considerations in mind, we introduce STMDiff, a two-stage network architecture for multiplexing the first scanned CT image to assist in the second PET attenuation correction. The pretrained multimodal spatiotemporal matching network is designed to find k-best-matched CT images while diffusion-based attenuation correction network takes identified Top-k CT images as a condition prior to synthesize AC PET images at second scanning time point.

In summary, we make the following contributions:

(1) STMDiff integrates contrastive learning to extract shared alignment and identifies optimal matching within latent space, which provides a direct one-to-one mapping between non-simultaneously acquired PET and CT data in DTPI enhancing its potential application.

(2) STMDiff introduces a dual-stage network architecture based on diffusion models with the identified k-best-matched CT images as prior conditions. Our method has successfully repurposed CT images while avoiding the alignment errors associated with rigid registration.

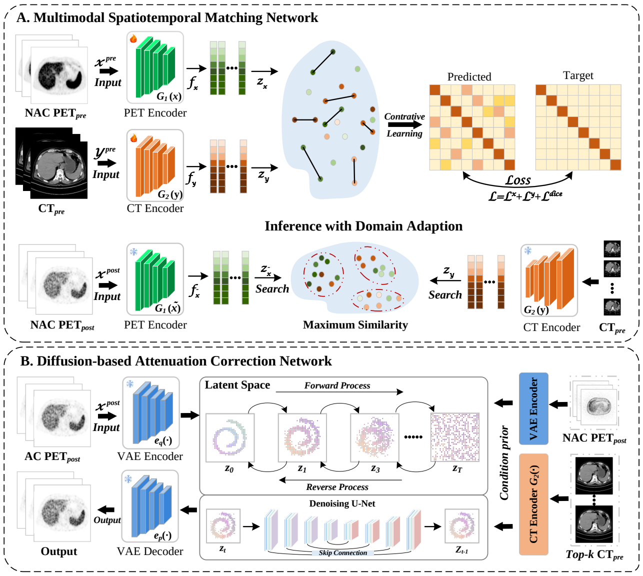

Fig. 2. Overall framework of the proposed STMDiff. A shows the network details of the multimodal spatiotemporal matching network, and B shows the details of diffusion-based attenuation correction network.

2 Methodology

Given datasets $\mathcal{D}_{\text{pre}}$ and $\mathcal{D}_{\text{post}}$ obtained in DTPI, where $\mathcal{D}_{\text{pre}}$ is the initial scanning data in DTPI while $\mathcal{D}_{\text{post}}$ contains the information gathered from the subsequent scan. Both datasets consist of aligned image pairs $\{x_i^d, y_i^d\}_{i=1}^N$, where N is the number of image pairs, d represents the datasets to which it belongs. x_i^d and y_i^d are images belonging to PET $\mathcal{X} \in \mathbb{R}^{C\times H\times W}$ and CT $\mathcal{Y} \in \mathbb{R}^{C\times H\times W}$, respectively. The image size is $C \times H \times W$, where H and W represent the height and width of images, $C = 1$. Fig. 2 illustrates the complete STMDiff framework. The individual components are displayed in detail below.

2.1 Multimodal Spatiotemporal Matching Network

We can observe that both scans in DTPI from the same patient shares many similar attributes, like geometry. The previous work proved that modal-independent information is useful in image matching [14]. Based on this, we map multimodal images into the shared space and then search for the optimal matching. Then, we introduce spatiotemporal contrastive learning, where images of the same anatomical structure correspond to similar representations, while those of different anatomical structures correspond to diverse features in Fig. 2(A).

The objective of matching network is to learn function $G_1(\cdot)$ and $G_2(\cdot)$ based on $\mathcal{D}_{\text{pre}}$, which enables paired images in domain $\mathcal{X}$ and $\mathcal{Y}$ to the latent space. In the multi-modal dataset $\mathcal{D}_{\text{pre}}$, the image pair $\{x_i^{pre}, y_j^{pre}\}_{i=j}$ is matched, $\{x_i^{pre}, y_j^{pre}\}_{i \neq j}$ are viewed as distinct scenes and their latent representation are negatives pairs. The extracted latent representations are defined as $z_x^i = G_1(x_i^{pre})$ and $z_y^i = G_2(x_i^{pre})$, where $G_1(\cdot)$ and $G_2(\cdot)$ are implemented with ResNet-19. The multimodal spatiotemporal matching network is trained to optimize the similarity among these latent representations. Specifically, it aims to pull together the representations of similar images $\{z_x^i, z_y^i\}_{i=j}$, and push apart the representations of dissimilar images $\{z_x^i, z_y^i\}_{i \neq j}$.

Considering the generalization of contrastive learning[15], STMDiff takes delayed PET images x_i^{post} and CT images y_j^{pre} as input in the inference stage. By comparing the similarity between their latent representations, the matching network identifies the k-best-matched CT images from first scan time point that correspond to the second scanned PET image, where k fixed as 5.

Loss function: We randomly sample some data $\{x_i^{pre}, y_j^{pre}\}$ at each training session and trained the network to optimize the functions $G_1(\cdot)$ and $G_2(\cdot)$. The InfoNCE loss function [16] for updating $G_1(\cdot)$ and $G_2(\cdot)$ is formulated as:

$$
\begin{aligned}
\mathcal{L}^{\mathcal{X}} &= -\sum_{i \in M} log \frac{exp(s(z_x^i, z_y^i)) \times \frac{1}{\tau}}{\sum_{i \notin M, i \neq j} exp(s(z_x^i, z_y^j)) \times \frac{1}{\tau} + exp(s(z_x^i, z_x^j)) \times \frac{1}{\tau}} \\
\mathcal{L}^{\mathcal{Y}} &= -\sum_{i \in M} log \frac{exp(s(z_x^i, z_y^i)) \times \frac{1}{\tau}}{\sum_{i \notin M, i \neq j} exp(s(z_x^i, z_y^j)) \times \frac{1}{\tau} + exp(s(z_y^i, z_y^j)) \times \frac{1}{\tau}}
\end{aligned}
\tag{1}
$$

where τ is a temperature coefficient to adjust the dynamic range and $\tau = 0.5$. M is the sampling subset for training. $L1$ norm is utilized as $s(\cdot)$ to measure the similarity of the extracted latent representations.

Due to the brief interval separating the two scans in DTPI, the geometrical consistency of PET/CT images from the same patient is maintained. Next, we apply morphological operations (opening, closing and threshold segmentation) for PET/CT images to accelerate optimal matching. Then, the Dice loss $\mathcal{L}^{dice}$ based on contour information for geometry constraint is introduced. In summary, the total loss function is defined as :

$$\mathcal{L} = \mathcal{L}^{\mathcal{X}} + \mathcal{L}^{\mathcal{Y}} + \mathcal{L}^{dice} \tag{2}$$

2.2 Diffusion-based Attenuation Correction Network

The main idea of diffusion models is to learn the target data distribution $q(x_0)$ (AC PET images in our case) using a neural network [17]. When the data distribution is learned, we can synthesize a new sample. Diffusion models contain the forward diffusion process and the reverse diffusion process: 1)The forward diffusion process gradually adds Gaussian noise to the image while progressively eroding the original details, eventually transforming the image into noise [18, 19]; 2) The reverse diffusion process then learns to reconstruct the original image by iteratively removing the noise, guided by the learned distribution, as shown in Fig. 2(B). The formulas of the diffusion-based model are as followed:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\gamma_t}x_0, (1-\gamma_t)I); \quad x_t = \sqrt{\gamma_t}x_0 + \sqrt{1-\gamma_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{3}$$

where $\gamma_t = \prod_{i=1}^{t} \alpha_i$. since x_0 is unknown during inference, the transition distribution $p_\theta(x_{t-1}|x_t)$ is used to approximate the reverse diffusion posterior:

$$\begin{aligned} p_\theta(x_{t-1}|x_t) &= \mathcal{N}\left(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I\right) \\ \mu_\theta(x_t, t) &= \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) \end{aligned} \tag{4}$$

Here, ϵ_θ denotes a neural network, and its training objective of $\epsilon_\theta(x_t, t)$ can be formulated as follows:

$$\mathbb{E}_{x,\epsilon,t\sim[1,T]} \left\|\epsilon - \epsilon_\theta(x_t, t)\right\|^2 . \tag{5}$$

Condition: Here we chose UNet as the ϵ_θ. To achieve PET attenuation correction instead of generating new samples, the network needs a input noisy counterparts as additional condition. Therefore, (x_t, t) changed to (x_t, t, x_{NAC}), where x_{NAC} denotes the input NAC PET images. Apart from handling x_{NAC} images, we also incorporate k-best-matched CT images as the condition. When processing CT and NAC PET images within the diffusion model, we employ two encoders: our pretrained embedder G_{ve} following the variational autoencoder (VAE) model derived from Stable Diffusion and the function $G_2(\cdot)$. The parameters are fixed during the network training process. The function $G_2(\cdot)$ maps CT images to the latent space to obtain their representations z_y. Moreover, G_{ve} processes the NAC PET images and generates their latent representations e_{nac}:

$$\begin{aligned} z_y &= G_2(y), \\ e_{nac} &= G_{ve}(x_{nac}), \\ c &= \text{concate}(z_y, e_{nac}). \end{aligned} \tag{6}$$

Finally, the feature fusion is conducted in the latent space through the most straightforward concatenation approach.

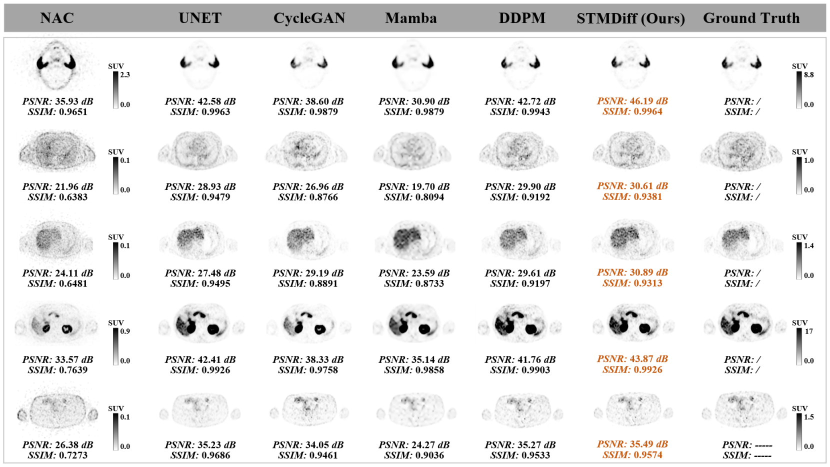
NAC
UNET
CycleGAN
Mamba
DDPM
STMDiff (Ours)
Ground Truth
PSNR: 35.93 dB SSIM: 0.9651
SUV 2.3 0.0
PSNR: 42.58 dB SSIM: 0.9963
PSNR: 38.60 dB SSIM: 0.9879
PSNR: 30.90 dB SSIM: 0.9879
PSNR: 42.72 dB SSIM: 0.9943
PSNR: 46.19 dB SSIM: 0.9964
PSNR: / SSIM: /
SUV 8.8 0.0
PSNR: 21.96 dB SSIM: 0.6383
SUV 0.1 0.0
PSNR: 28.93 dB SSIM: 0.9479
PSNR: 26.96 dB SSIM: 0.8766
PSNR: 19.70 dB SSIM: 0.8094
PSNR: 29.90 dB SSIM: 0.9192
PSNR: 30.61 dB SSIM: 0.9381
PSNR: / SSIM: /
SUV 1.0 0.0
PSNR: 24.11 dB SSIM: 0.6481
SUV 0.1 0.0
PSNR: 27.48 dB SSIM: 0.9495
PSNR: 29.19 dB SSIM: 0.8891
PSNR: 23.59 dB SSIM: 0.8733
PSNR: 29.61 dB SSIM: 0.9197
PSNR: 30.89 dB SSIM: 0.9313
PSNR: / SSIM: /
SUV 1.4 0.0
PSNR: 33.57 dB SSIM: 0.7639
SUV 0.9 0.0
PSNR: 42.41 dB SSIM: 0.9926
PSNR: 38.33 dB SSIM: 0.9758
PSNR: 35.14 dB SSIM: 0.9858
PSNR: 41.76 dB SSIM: 0.9903
PSNR: 43.87 dB SSIM: 0.9926
PSNR: / SSIM: /
SUV 17 0.0
PSNR: 26.38 dB SSIM: 0.7273
SUV 0.1 0.0
PSNR: 35.23 dB SSIM: 0.9686
PSNR: 34.05 dB SSIM: 0.9461
PSNR: 24.27 dB SSIM: 0.9036
PSNR: 35.27 dB SSIM: 0.9533
PSNR: 35.49 dB SSIM: 0.9574
PSNR: ----- SSIM: -----
SUV 1.5 0.0

Fig. 3. Visualizations of generated results for different structures (Brain, Lungs, Liver, Kidney and Pelvis) with their corresponding PSNR and SSIM values. The best results are highlighted in orange.

3 Experiments

Models and Hyperparameters: 1)For Multimodal Spatiotemporal Matching Network: adaptive moment estimation (ADAM) optimizer was used to minimize the loss $\mathcal{L}$. The batch size was 128 with 300 epochs trained. The learning rate was initially set at 1 e^{-4} and halved after every 100 epochs; 2)For Diffusion-based Attenuation Correction Network: we used an exponential moving average (EMA) with a decay rate of 0.999 [25]. The Adam optimizer was utilized with the a learning rate of 1 e^{-4}. We downloaded the Variational Autoencoder (VAE) model from Stable Diffusion (Factor:8) and fine-tuned the VAE model on our datasets. All experiments were performed within the latent space generated by the VAE encoder, and the sampling results were subsequently reconstructed into origin form using the VAE decoder[24]. The linear variance schedule t had a maximum value of 1000. Ordinary Differential Equation (ODE) sampling was used for our experiments. All the experiments were conducted on a NVIDIA GeForce RTX 4090 with Ubuntu 24.04 LTS system.

Datasets and Preprocessing: All PET/CT images were acquired using the uEXPLORER total-body PET/CT scanner. The study included 104 patients underwent dual-time-point total-body PET/CT imaging. 84 subjects (56,532 slices) were selected for training, 10 subjects (6,730 slices) for validation, and 10 subjects (6,730 slices) for testing. The initial scanning time was about 60 minutes after ^{68}Ga-prostate-specific membrane antigen-11 (^{68}Ga-PSMA) injection and the second time was around 150 minutes. PET image are with resolution of 160×160 and CT images are 256×256. We resample the CT images to match

the dimensions of the PET images. Before model training, DICOM-format PET images were transformed to standardized uptake value (SUV) units.
Evaluation: Our comparison methods include UNET [20], CycleGAN [21], Mamba [22], DDPM[23].To assess the quality of the generated AC PET images, three common metrics were used for analyses: the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) and root mean squared Error (RMSE). In addition to objectively assessing image quality with quantitative indicators, we performed regoin of interest (ROI) analysis on different hypermetabolic regions to compare distribution consistency.

4 Results

Qualitative Results: Fig. 3 shows the qualitative results obtained from STMDiff and other comparison methods. The generated AC PET images by STMDiff are visually close to the ground truth (GT), demonstrating the effective image noise suppression effect. To quantitatively compare the results generated by different methods, we provide the corresponding PSNR and SSIM values. Compared to NAC images, all methods achieve significant improvement on both quantitative results and image quality. Among them, our proposed method obtain better model performance with highest PSNR values (PSNR > 30.00 dB).
Clinical Assessment: In addition to the qualitative image analysis, we perform a consistency analysis for the SUV distributions of the PET images, considering the liver and kidney as ROIs, as shown in Fig. 4. The median and upper and lower quartile scores of our results are approximately identical to those of the GT as well as the shape of the violin plots. To evaluate distribution similarity, we calculate the Pearson correlation coefficients (PCC) for different methods within ROIs. STMDiff achieves highest PCC values (PCC > 0.96) in both liver and kidney regions, which is beneficial to study metabolism. Moreover, the T-test shows that the difference is statistically significant (p < 0.001).

Table 1. Quantitative results (Mean ± Std) on test data for different methods in terms of PSNR, SSIM and RMSE. The best results are indicated in bold.

Methods	PSNR ↑	SSIM ↑	RMSE ↓
NAC	30.07±6.90	0.83±0.11	3.50±1.96
UNET [20]	36.55±7.41*	0.89±0.25*	2.88±2.06*
CycleGAN [21]	35.05±5.33*	0.95±0.04*	2.46±1.25*
Mamba [22]	33.34±6.92*	0.95±0.05*	2.66±1.17*
DDPM [23]	36.45±6.13*	0.95±0.04*	2.46±1.04*
STMDiff	**37.71±6.86***	**0.96±0.04***	**2.35±1.03***

* denotes $p < 0.05$, corresponding to a significant difference.

Quantitative Results: Table. 1 reports quantitative metrics between GT and synthesized AC PET for all methods. Compared with other baseline methods, the

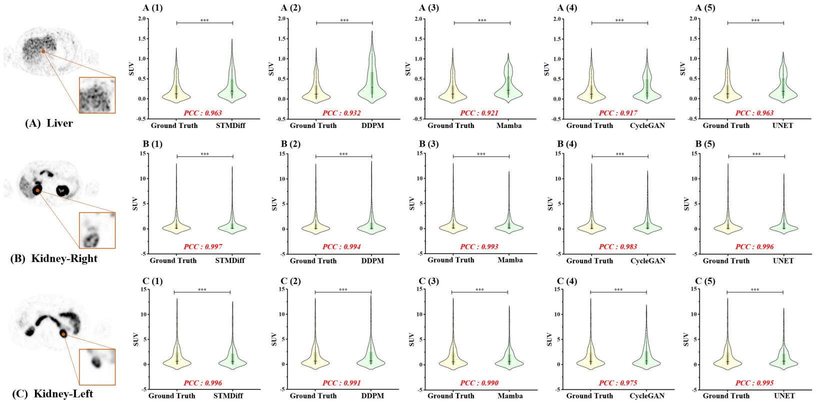

Fig. 4. Violin plots for generated AC PET images obtained by all methods within liver and kidney. "PCC" denotes the Pearson correlation coefficient. "$*\,*\,*$" denotes that the p value is less than 0.01.

STMDiff generates AC PET with the highest quality, owning the lowest RMSE (2.35 ± 1.03), and highest PSNR (37.72 ± 6.85 dB), SSIM (0.96 ± 0.03). Since the comparison methods directly generate AC PET images, the UNET model also achieves good performance. While the Mamba method produces blurred images, which may be caused by inadequate parameter training.

Ablation Study: Experimental results demonstrate the effectiveness of the proposed method. We conduct ablation studies to demonstrate the effectiveness of incorporating k-best-matched CT image priors. The baseline model utilize only NAC images as priors to generate AC PET images. In contrast, STMDiff additionally incorporated the k-best-matched CT images. The corresponding results are presented in Fig. 5. And the quantitative results indicates that CT prior significantly elevates generation quality.

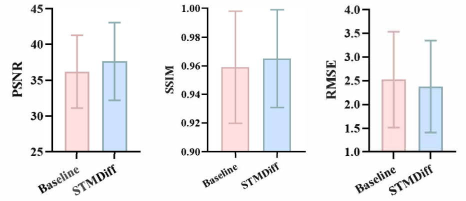

Fig. 5. Quantitative results for ablation study in terms of PSNR, SSIM and RMSE.

In our initial investigation, parameter k was set to 5. To further examine the impact of the number of parameters k integrated into the diffusion model, a second ablation study was conducted. We retrained the model with 1 to 5 CT image priors and quantitatively evaluated the results. The results indicate that the optimal quantitative metrics were achieved when $k = 2$ (PSNR: 38.19 dB, SSIM: 0.97, RMSE:2.25). Compared with those attained when $k = 5$, the PSNR increased by 2.46%, the SSIM increased by 1.00%, the RMSE decreased by 5.86%. This suggests that increasing the number of CT priors can enhance the performance of the model, with the best results obtained at $k = 2$. Therefore, we plan to further optimize the experimental results based on the optimal parameters in future work to ensure the accuracy and effectiveness of our research.

5 Conclusion

We develop a STMDiff for PET attenuation correction in DTPI to eliminate the radiation risks of repeated CT scan. The core components of STMDiff consist of two networks: 1) The multimodal spatiotemporal matching network successfully searched k-best-matched CT images from initial scan; 2) The diffusion-based attenuation correction network takes identified Top-k CT images as a condition prior to synthesize AC PET images at second scanning time point. Compared to the state-of-art methods, the STMDiff excels in suppressing image noise (PSNR: 37.72 ± 6.85 dB; SSIM: 0.96 ± 0.03; RMSE: 2.35 ± 1.03) and retaining good consistency of anatomical structures (PCC > 0.96) on test dataset, which shows great potential in reducing additional radiation hazard of repeated scans.

Acknowledgments. This work was supported by the National Key Research and Development Program of China (2024YFE0202400), National Natural Science Foundation of China (12326607, 82372038), the Natural Science Foundation of Guangdong Province in China (2023B1515120007, 2024B1515040018), the Shenzhen Excellent Technological Innovation Talent Training Project of China (RCJC20200714114436080), the Shenzhen Science and Technology Program of China (KJZD20240903101307010, JCYJ20220818101804009, SGDX20230821092359002), the Shenzhen Medical Research Fund of China (B2301002).

Disclosure of Interests. Yihan Wang is the employee of Central Research Institute, United Imaging Healthcare Group. Other authors have no competing interests to declare that are relevant to the content of this article.

References

1. Z. Huang, W. Li, Y. Wu, N. Guo, L. Yang, N. Zhang, Z. Pang, Y. Yang, Y. Zhou, Y. Shang, H. Zheng, D. Liang, M. Wang, and Z. Hu, "Short-axis pet image quality improvement based on a uexplorer total-body pet system through deep learning," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 1, pp. 27–39, 2023.
2. Y. Wang, L. Dong, H. Zhao, L. Li, G. Huang, W. Xue, J. Liu, and R. Chen, "The superior detection rate of total-body [68ga]ga-psma-11 pet/ct compared to short axial field-of-view [68ga]ga-psma-11 pet/ct for early recurrent prostate cancer patients with psa <0.2 ng/ml after radical prostatectomy," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 8, pp. 2484–2494, 2024.

3. Y. Wu, F. Fu, N. Meng, Z. Wang, X. Li, Y. Bai, Y. Zhou, D. Liang, H. Zheng, Y. Yang, M. Wang, and T. Sun, “The role of dynamic, static, and delayed total-body pet imaging in the detection and differential diagnosis of oncological lesions,” *Cancer Imaging*, vol. 24, no. 1, p. 2, 2024.
4. X. Li, J. M. Johnson, R. M. Strigel, L. C. H. Bancroft, S. A. Hurley, S. I. Z. Estakhraji, M. Kumar, A. M. Fowler, and A. B. McMillan, “Attenuation correction and truncation completion for breast pet/mr imaging using deep learning,” *Physics in Medicine and Biology*, vol. 69, no. 4, 2024.
5. Y. Lu, F. Kang, D. Zhang, Y. Li, H. Liu, C. Sun, H. Zeng, L. Shi, Y. Zhao, and J. Wang, “Deep learning-aided respiratory motion compensation in pet/ct: addressing motion induced resolution loss, attenuation correction artifacts and pet-ct misalignment,” *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 52, no. 1, pp. 62–73, 2024.
6. D. Bau, J. Y. Zhu, J. Wulff, W. Peebles, H. Strobelt, B. L. Zhou, and A. Torralba, “Seeing what a gan cannot generate,” *2019 IEEE/Cvf International Conference on Computer Vision (Iccv 2019)*, pp. 4501–4510, 2019.
7. Y. T. Li, I. Yakushev, D. M. Hedderich, and C. Wachinger, “Pasta: Pathology-aware mri to pet cross-modal translation with diffusion models,” *Medical Image Computing and Computer Assisted Intervention - Miccai 2024, Pt Vii*, vol. 15007, pp. 529–540, 2024.
8. R. Guo, S. Xue, J. Hu, H. Sari, C. Mingels, K. Zeimpekis, G. Prenosil, Y. Wang, Y. Zhang, M. Viscione, R. Sznitman, A. Rominger, B. Li, and K. Shi, “Using domain knowledge for robust and generalizable deep learning-based ct-free pet attenuation and scatter correction,” *Nature Communications*, vol. 13, no. 1, p. 5882, 2022.
9. W. Li, Z. Huang, Z. Chen, Y. Jiang, C. Zhou, X. Zhang, W. Fan, Y. Zhao, L. Zhang, L. Wan, Y. Yang, H. Zheng, D. Liang, and Z. Hu, “Learning ct-free attenuation-corrected total-body pet images through deep learning,” *European Radiology*, vol. 34, no. 9, pp. 5578–5587, 2024.
10. H. Wang, Y. Wang, Q. Xue, Y. Zhang, X. Qiao, Z. Lin, J. Zheng, Z. Zhang, Y. Yang, M. Zhang, Q. Huang, Y. Huang, T. Cao, J. Wang, and B. Li, “Optimizing mr-based attenuation correction in hybrid pet/mr using deep learning: validation with a flatbed insert and consistent patient positioning,” *European Journal of Nuclear Medicine and Molecular Imaging*, 2025.
11. M. Ajith and V. D. Calhoun, “Conditional denoising diffusion probabilistic models with attention for subject-specific brain network synthesis,” *bioRxiv*, 2025.
12. F. Khader, G. Muller-Franzes, S. Tayebi Arasteh, T. Han, C. Haarburger, M. Schulze-Hagen, P. Schad, S. Engelhardt, B. Baessler, S. Foersch, J. Stegmaier, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn, “Denoising diffusion probabilistic models for 3d medical image generation,” *Scientific Reports*, vol. 13, no. 1, p. 7303, 2023.
13. Y. Li, J. Guo, H. Qiu, F. Chen, and J. Zhang, “Denoising diffusion probabilistic models and transfer learning for citrus disease diagnosis,” *Frontiers in Plant Science*, vol. 14, p. 1267810, 2023.
14. H. Xu, J. Yuan, and J. Ma, “Murf: Mutually reinforcing multi-modal image registration and fusion,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 148–12 166, 2023.
15. W. Huang, M. Yi, X. Zhao, and Z. Jiang, “Towards the generalization of contrastive self-supervised learning,” *arXiv preprint arXiv:2111.00743*, 2021.
16. 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," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

17. A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *arXiv:2102.09672*, 2021.
18. J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv:2010.02502*, 2020.
19. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
20. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *Medical Image Computing and Computer-Assisted Intervention, Pt Iii*, vol. 9351, pp. 234–241, 2015.
21. J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, Conference Proceedings, pp. 2223–2232.
22. Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie, "Sigma: Siamese mamba network for multi-modal semantic segmentation," *arXiv preprint arXiv:2404.04256*, 2024.
23. R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, Conference Proceedings, pp. 10 684–10 695.
24. Z. Wang, L. Zhang, L. Wang, and Z. Zhang, "Soft masked mamba diffusion model for ct to mri conversion," *arXiv preprint arXiv:2406.15910*, 2024.
25. A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.