PET-Diffusion: Unsupervised PET Enhancement Based on the Latent Diffusion Model

Caiwen Jiang[1], Yongsheng Pan[1], Mianxin Liu[3], Lei Ma[1], Xiao Zhang[1], Jiameng Liu[1], Xiaosong Xiong[1], and Dinggang Shen[1,2,4(✉)]

[1] School of Biomedical Engineering, ShanghaiTech University, Shanghai, China
{jiangcw,panysh,dgshen}@shanghaitech.edu.cn
[2] Shanghai United Imaging Intelligence Co., Ltd., Shanghai 200230, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[4] Shanghai Clinical Research and Trial Center, Shanghai 201210, China

Abstract. Positron emission tomography (PET) is an advanced nuclear imaging technique with an irreplaceable role in neurology and oncology studies, but its accessibility is often limited by the radiation hazards inherent in imaging. To address this dilemma, PET enhancement methods have been developed by improving the quality of low-dose PET (LPET) images to standard-dose PET (SPET) images. However, previous PET enhancement methods rely heavily on the paired LPET and SPET data which are rare in clinic. Thus, in this paper, we propose an unsupervised PET enhancement (uPETe) framework based on the latent diffusion model, which can be trained only on SPET data. Specifically, our SPET-only uPETe consists of an encoder to compress the input SPET/LPET images into latent representations, a latent diffusion model to learn/estimate the distribution of SPET latent representations, and a decoder to recover the latent representations into SPET images. Moreover, from the theory of actual PET imaging, we improve the latent diffusion model of uPETe by 1) adopting **PET image compression** for reducing the computational cost of diffusion model, 2) using **Poisson diffusion** to replace Gaussian diffusion for making the perturbed samples closer to the actual noisy PET, and 3) designing **CT-guided cross-attention** for incorporating additional CT images into the inverse process to aid the recovery of structural details in PET. With extensive experimental validation, our uPETe can achieve superior performance over state-of-the-art methods, and shows stronger generalizability to the dose changes of PET imaging. The code of our implementation is available at https://github.com/jiang-cw/PET-diffusion.

Keywords: Positron emission tomography (PET) · Enhancement · Latent diffusion model · Poisson diffusion · CT-guided cross-attention

1 Introduction

Positron emission tomography (PET) is a sensitive nuclear imaging technique, and plays an essential role in early disease diagnosis, such as cancers and

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14220, pp. 3–12, 2023.
https://doi.org/10.1007/978-3-031-43907-0_1

Alzheimer's disease [8]. However, acquiring high-quality PET images requires injecting a sufficient dose (standard dose) of radionuclides into the human body, which poses unacceptable radiation hazards for pregnant women and infants even following the As Low As Reasonably Achievable (ALARA) principle [19]. To reduce the radiation hazards, besides upgrading imaging hardware, designing advanced PET enhancement algorithms for improving the quality of low-dose PET (LPET) images to standard-dose PET (SPET) images is a promising alternative.

In recent years, many enhancement algorithms have been proposed to improve PET image quality. Among the earliest are filtering-based methods such as non-local mean (NLM) filter [1], block-matching 3D filter [4], bilateral filter [7], and guided filter [22], which are quite robust but tend to over-smooth images and suppress the high-frequency details. Subsequently, with the development of deep learning, the end-to-end PET enhancement networks [9,14,21] were proposed and achieved significant performance improvement. But these supervised methods relied heavily on the paired LPET and SPET data that are rare in actual clinic due to radiation exposure and involuntary motions (e.g., respiratory and muscle relaxation). Consequently, unsupervised PET enhancement methods such as deep image prior [3], Noise2Noise [12,20], and their variants [17] were developed to overcome this limitation. However, these methods still require LPET to train models, which contradicts with the fact that only SPET scans are conducted in clinic.

Fortunately, the recent glowing diffusion model [6] provides us with the idea for proposing a clinically-applicable PET enhancement approach, whose training only relies on SPET data. Generally, the diffusion model consists of two reversible processes, where the forward diffusion adds noise to a clean image until it becomes pure noise, while the reverse process removes noise from pure noise until the clean image is recovered. By combining the mechanics of diffusion model with the observation that the main differences between LPET and SPET are manifested as levels of noises in the image [11], we can view LPET and SPET as results at different stages in an integrated diffusion process. Therefore, when a diffusion model (trained only on SPET) can recover noisy samples to SPET, this model can also recover LPET to SPET. However, extending the diffusion model developed for 2D photographic images to PET enhancement still faces two problems: a) three-dimensionsal (3D) PET images will dramatically increase the computational cost of diffusion model; b) PET is the detail-sensitive images and may be introduced/lost some details during the procedure of adding/removing noise, which will affect the downstream diagnosis.

Taking all into consideration, we propose the SPET-only unsupervised PET enhancement (uPETe) framework based on the latent diffusion model. Specifically, uPETe has an encoder-<diffusion model>-decoder structure that first uses the encoder to compress input the LPET/SPET images into latent representations, then uses the latent diffusion model to learn/estimate the distribution of SPET latent representations, and finally uses the decoder to recover SPET images from the estimated SPET latent representations. The keys of our uPETe

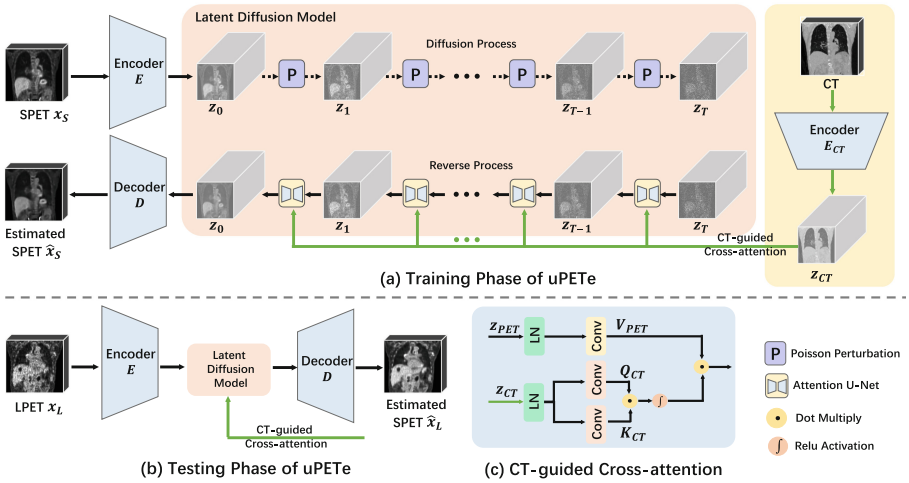

Fig. 1. Overview of proposed uPETe. (a) and (b) provide the framework of uPETe as well as depict its implementation during both the training and testing phases, and (c) illustrates the details of CT-guided cross-attention.

include 1) compressing the 3D PET images into a lower dimensional space for reducing the computational cost of diffusion model, 2) adopting the Poisson noise, which is the dominant noise in PET imaging [20], to replace the Gaussian noise in the diffusion process for avoiding the introduction of details that are not existing in PET images, and 3) designing CT-guided cross-attention to incorporate additional CT images into the inverse process for helping the recovery of structural details in PET.

Our work had three main features/contributions: i) proposing a clinically-applicable unsupervised PET enhancement framework, ii) designing three targeted strategies for improving the diffusion model, including PET image compression, Poisson diffusion, and CT-guided cross-attention, and iii) achieving better performance than state-of-the-art methods on the collected PET datasets.

2 Method

The framework of uPETe is illustrated in Fig. 1. When given an input PET image x (i.e., SPET for training and LPET for testing), x is first compressed into the latent representation z_0 by the encoder E. Subsequently, z_0 is fed into a latent diffusion model followed by the decoder D to output the expected SPET image $\hat{x}$. In addition, a specialized encoder E_{CT} is used to compress the CT image corresponding to the input PET image into the latent representation z_{CT}, which is fed into each denoising network for CT-guided cross-attention. In the following, we introduce the details of image compression, latent diffusion model, and implementation.

2.1 Image Compression

The conventional diffusion model is computationally-demanding due to its numerous inverse denoising steps, which severely restricts its application to 3D PET enhancement. To overcome this limitation, we adopt two strategies including 1) compressing the input image and 2) reducing the diffusion steps (as described in Sect. 2.3).

Similar to [10,18], we adopt an autoencoder (E and D) to compress the 3D PET images into a lower dimensional but more compact space. The crucial aspects of this process is to ensure that the latent representation contains the necessary and representative information for the input image. To achieve this, we train the autoencoder by a combination of perceptual loss [24] and patch-based adversarial loss [5], instead of simple voxel-level loss such as L_2 or L_1 loss. Among them, the perceptual loss, designed on a pre-trained 3D ResNet [2], constrains higher-level information such as texture and semantic content, and the patch-based adversarial loss ensures globally coherent while remaining locally realistic. Let $x \in \mathbb{R}^{H,W,Z}$ denote the input image and $z_0 \in \mathbb{R}^{h,w,z,c}$ denote the latent representation. The compression process can be formulated as $\hat{x} = D(z_0) = D(E(x))$. In this way, we compress the input image by a factor of $f = H/h = W/w = Z/z$. The results of SPET estimation under different compression rates f are provided in the supplement.

2.2 Latent Diffusion Model

After compressing the input PET image, its latent representation is fed into the latent diffusion model, which is the key to achieving the SPET-only unsupervised PET enhancement. As described above, the LPET can be viewed as noisy SPET (even in the compressed space), so the diffusion process from SPET to pure noise actually covers the situations of LPET. That is, the diffusion model trained with SPET is capable of estimating SPET from the noisy sample (diffused from LPET). But the diffusion model is developed from photographic images, which have significant difference with the detail-sensitive PET images. To improve its applicability for PET images, we design several targeted strategies for the diffusion process and inverse process, namely *Poisson diffusion* and *CT-guided cross-attention*, respectively.

Poisson Diffusion. In conventional diffusion models, the forward process typically employs Gaussian noise to gradually perturb input samples. However, in PET images, the dominant source of noise is Poisson noise, rather than Gaussian noise. Considering this, in our uPETe we choose to adopt Poisson diffusion to perturb the input samples, which facilitates the diffusion model for achieving better performance on the PET enhancement task.

Let z_t be the perturbation sample in Poisson diffusion, where $t = 0, 1, ..., T$. Then the Poisson diffusion can be formulate as follows:

$$z_t = perturb(z_{t-1}, \lambda_t), \quad \lambda_1 < \lambda_2 < ... < \lambda_T. \tag{1}$$

At each diffusion step, we apply the *perturb* function to the previous perturbed sample z_{t-1} by imposing a Poisson noise with an expectation of λ_t, which is linearly interpolated from $[0, 1]$ and incremented with t. In our implementation, we apply the same Poisson noise imposition operation as in [20], i.e., applying Poisson deviates on the projected sinograms, to generate a sequence of perturbed samples with increasing Poisson noise intensity as the step number t increases.

CT-Guided Cross-Attention. The attenuation correction of PET typically relies on the corresponding anatomical image (CT or MR), resulting in a PET scan usually accompanied by a CT or MR scan. To fully utilize the extra-modality images (i.e., CT in our work) as well as improve the applicability of diffusion models, we design a CT-guided cross-attention to incorporate the CT images into the reverse process for assisting the recovery of structural details.

As shown in Fig. 1, to achieve a particular SPET estimation, the corresponding CT image is first compressed into the latent representation $z_{_{CT}}$ by encoder E_{CT}. Then $z_{_{CT}}$ is fed into a denoising attention U-Net [16] at each step for calculation of cross-attention, where the query Q and key K are calculated from $z_{_{CT}}$ while the value V is still calculated from the output of the previous layer because our final goal is SPET estimation. Denoting the output of previous layer as $z_{_{PET}}$, the CT-guided cross-attention can be formulated as follows:

$$\begin{aligned} Output &= softmax(\frac{Q_{CT}K_{CT}^{T}}{\sqrt{d}} + B) \cdot V_{PET}, \\ Q_{CT} = Conv_Q(z_{_{CT}}), \quad K_{CT} &= Conv_K(z_{_{CT}}), \quad V_{PET} = Conv_V(z_{_{PET}}), \end{aligned} \tag{2}$$

where d is the number of channels, B is the position bias, and $Conv(\cdot)$ denotes the $1 \times 1 \times 1$ convolution with stride of 1.

2.3 Implementation Details

Typically, the trained diffusion model generates target images from random noise, requiring a large number of steps T to make the final perturbed sample ($z_{_T}$) close to pure noise. However, in our task, the target SPET image is generated from a given LPET image during testing, and making $z_{_T}$ as close to pure noise as possible is not necessary since the remaining PET-related information can also benefit the image recovery. Therefore, we can considerably reduce the number of diffusion steps T to accelerate the model training, and T is set to 400 in our implementation. We evaluate the quantitative results using two metrics, including Peak Signal to Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM).

Table 1. Quantitative results of ablation analysis, in terms of PSNR and SSIM.

Method	PSNR [dB]↑	SSIM ↑
LDM	23.732 ± 1.264	0.986 ± 0.010
LDM-P	24.125 ± 1.072	0.987 ± 0.009
LDM-CT	25.348 ± 0.822	0.990 ± 0.006
LDM-P-CT	**25.817 ± 0.675**	**0.992 ± 0.004**

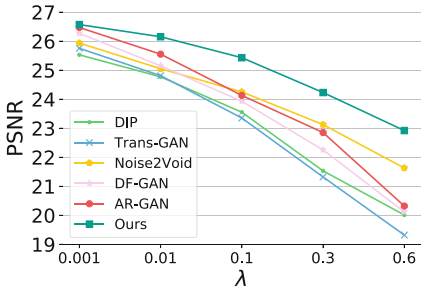

Fig. 2. Generalizability to dose changes.

3 Experiments

3.1 Dataset

Our dataset consists of 100 SPET images for training and 30 paired LPET and SPET images for testing. Among them, 50 chest-abdomen SPET images are collected from (total-body) uEXPLORER PET/CT scanner [25], and 20 paired chest-abdomen images are collected by list mode of the scanner with 256 MBq of [^{18}F]-FDG injection. Specifically, the SPET images are reconstructed by using the 1200 s data between 60–80 min after tracer injection, while the corresponding LPET images are simultaneously reconstructed by 120 s data uniformly sampled from 1200 s data.

As a basic data preprocessing, all images are resampled to voxel spacing of $2 \times 2 \times 2$ mm^3 and resolution of $256 \times 256 \times 160$, while their intensity range is normalized to $[0, 1]$ by min-max normalization. For increasing the training samples and reducing the dependence on GPU memory, we extract the overlapped patches of size $96 \times 96 \times 96$ from every whole PET image.

3.2 Ablation Analysis

To verify the effectiveness of our proposed strategies, i.e. *Poisson diffusion process* and *CT-guided cross-attention*, we design another four variant latent diffusion models (LDMs) with the same compression model, including: 1) LDM: standard LDM; 2) LDM-P: LDM with Poisson diffusion process; 3) LDM-CT: LDM with CT-guided cross-attention; 4) LDM-P-CT: LDM with Poisson diffusion process and CT-guided cross-attention. All methods use the same experimental settings, and their quantitative results are given in Table 1.

From Table 1, we can have the following observations. (1) LDM-P achieves better performance than LDM. This proves that the Poisson diffusion is more appropriate than the Gaussian diffusion for PET enhancement. (2) LDM-CT with the corresponding CT image for assisting denoising achieves better results than LDM. This can be reasonable as the CT image can provide anatomical information, thus benefiting the recovery of structural details (e.g., organ boundaries) in SPET images. (3) LDM-P-CT achieves better results than all other variants

Table 2. Quantitative comparison of our uPETe with several state-of-the-art PET enhancement methods, in terms of PSNR and SSIM, where * denotes unsupervised method and † denotes fully-supervised method.

Method	PSNR [dB]↑	SSIM ↑
DIP* [3]	22.538 ± 2.136	0.981 ± 0.015
Noisier2Noise * [23]	22.932 ± 1.983	0.983 ± 0.014
LA-GAN † [21]	23.351 ± 1.725	0.984 ± 0.012
MR-GDD * [17]	23.628 ± 1.655	0.985 ± 0.011
Trans-GAN† [13]	23.852 ± 1.522	0.985 ± 0.009
Noise2Void * [20]	24.263 ± 1.351	0.987 ± 0.009
DF-GAN† [9]	24.821 ± 0.975	0.989 ± 0.007
AR-GAN† [14]	25.217 ± 0.853	0.990 ± 0.006
uPETe*	**25.817 ± 0.675**	**0.992 ± 0.004**

on both PSNR and SSIM, which shows both of our proposed strategies contribute to the final performance. These three comparisons conjointly verify the effective design of our proposed uPETe, where the *Poisson diffusion process* and *CT-guided cross-attention* both benefit the PET enhancement.

3.3 Comparison with State-of-the-Art Methods

We further compare our uPETe with several state-of-the-art PET enhancement methods, which can be divided into two classes: 1) fully-supervised methods, including LA-GAN [21], Transformer-GAN (Trans-GAN) [13], Dual-frequency GAN (DF-GAN) [9], and AR-GAN [14]; 2) unsupervised methods, including deep image prior (DIP) [3], Noisier2Noise [23], magnetic resonance guided deep decoder (MR-GDD) [17], and Noise2Void [20]. The quantitative and qualitative results are provided in Table 2 and Fig. 3, respectively.

Quantitative Comparison: Table 2 shows that our uPETe outperforms all competing methods. Compared to the fully-supervised method AR-GAN which achieves sub-optimal performance, our uPETe does not require paired LPET and SPET, yet still achieves improvement. Additionally, uPETe also achieves noticeable performance improvement to Noise2Void (which is a supervised method). Specifically, the average improvement in PSNR and SSIM on SPET estimation are 1.554 dB and 0.005, respectively. This suggests that our uPETe can generate promising results without relying on paired data, demonstrating its potential for clinical applications.

Qualitative Comparison: In Fig. 3, we provide a visual comparison of SPET estimation for two typical cases. First, compared to unsupervised methods such as DIP and Noise2Void, the SPET images estimated by our uPETe have less noise but clearer boundaries. Second, our uPETe performs better on the structural details compared to the fully-supervised methods, i.e., missing unclear tissue (Trans-GAN) or introducing non-existing artifacts in PET image (DF-GAN).

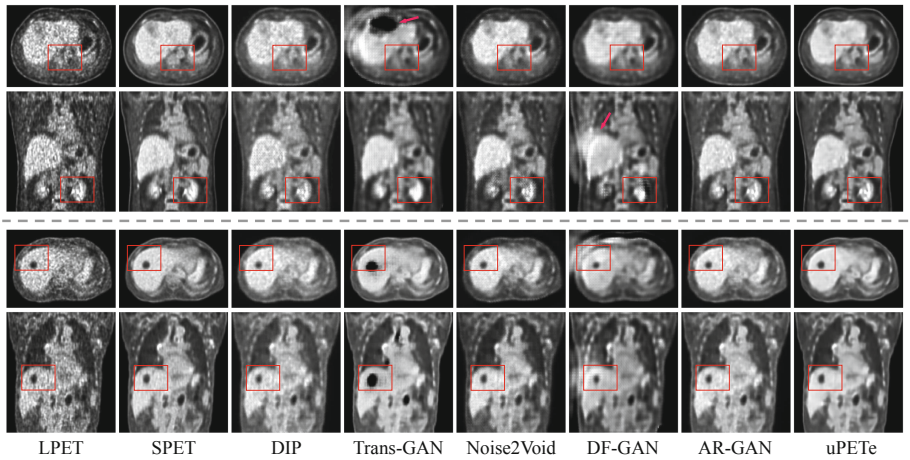

Fig. 3. Visual comparison of estimated SPET images on two typical cases. In each case, the first and second rows show the axial and coronal views, respectively, and from left to right are the input (LPET), ground truth (SPET), results by five other methods (3rd–7th columns), and the result by our uPETe (last column). Red boxes and arrows show areas for detailed comparison. (Color figure online)

Overall, these pieces of evidence demonstrate the superiority of our uPETe over state-of-the-art methods.

3.4 Generalization Evaluation

We further evaluate the generalizability of our uPETe to tracer dose changes by simulating Poisson noise on SPET to produce different doses for LPET, which is a common way to generate noisy PET data [20]. Notably, we do not need to retrain the models since they have been trained in Sect. 3.3. The quantitative results of our uPETe and five state-of-the-art methods are provided in Fig. 2.

As shown in Fig. 2, our uPETe outperforms the other five methods at all doses and exhibits a lower PSNR descent slope as dose decreases (i.e., λ increases), demonstrating its superior generalizability to dose changes. This is because uPETe is based on diffusion model, which simplifies the complex distribution prediction task into a series of simple denoising tasks and thus has strong generalizability. Moreover, we also find that the unsupervised methods (i.e., uPETe, Noise2Void, and DIP) have stronger generalizability than fully-supervised methods (i.e., AR-GAN, DF-GAN, and Trans-GAN) as they have a smoother descent slope. The main reason is that the unsupervised learning has the ability to extract patterns and features from the data based on the inherent structure and distribution of the data itself [15].

4 Conclusion and Limitations

In this paper, we have developed a clinically-applicable unsupervised PET enhancement framework based on the latent diffusion model, which uses only the clinically-available SPET data for training. Meanwhile, we adopt three strategies to improve the applicability of diffusion models developed from photographic images to PET enhancement, including 1) compressing the size of the input image, 2) using Poisson diffusion, instead of Gaussian diffusion, and 3) designing CT-guided cross-attention to enable additional anatomical images (e.g., CT) to aid the recovery of structural details in PET. Validated by extensive experiments, our uPETe achieved better performance than both state-of-the-art unsupervised and fully-supervised PET enhancement methods, and showed stronger generalizability to the tracer dose changes.

Despite the advance of uPETe, our current work still suffers from a few limitations such as (1) lacking theoretical support for our Poisson diffusion, which is just an engineering attempt, and 2) only validating the generalizability of uPETe on a simulated dataset. In our future work, we will complete the design of Poisson diffusion from theoretical perspective, and collect more real PET datasets (e.g., head datasets) to comprehensively validate the generalizability of our uPETe.

Acknowledgment. This work was supported in part by National Natural Science Foundation of China (No. 62131015), Science and Technology Commission of Shanghai Municipality (STCSM) (No. 21010502600), The Key R&D Program of Guangdong Province, China (No. 2021B0101420006), and the China Postdoctoral Science Foundation (Nos. BX2021333, 2021M703340).

References

1. Buades, A., Coll, B., Morel, J.: A non-local algorithm for image denoising. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 60–65 (2005)
2. Chen, S., Ma, K., Zheng, Y.: Med3D: transfer learning for 3D medical image analysis. arXiv preprint arXiv:1904.00625 (2019)
3. Cui, J., et al.: PET image denoising using unsupervised deep learning. Eur. J. Nucl. Med. Mol. Imaging **46**(13), 2780–2789 (2019)
4. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising with block-matching and 3D filtering. Image Process. Algorithms Syst. Neural Netw. Mach. Learn. **6064**, 354–365 (2006)
5. Dosovitskiy, A., Brox, T.: Generating images with perceptual similarity metrics based on deep networks. In: Advances in Neural Information Processing Systems, vol. 29 (2016)
6. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems, vol. 33, pp. 6840–6851 (2020)
7. Hofheinz, F., et al.: Suitability of bilateral filtering for edge-preserving noise reduction in PET. EJNMMI Res. **1**(1), 1–9 (2011)

8. Jiang, C., Pan, Y., Cui, Z., Nie, D., Shen, D.: Semi-supervised standard-dose PET image generation via region-adaptive normalization and structural consistency constraint. IEEE Trans. Med. Imaging (2023)
9. Jiang, C., Pan, Y., Cui, Z., Shen, D.: Reconstruction of standard-dose PET from low-dose PET via dual-frequency supervision and global aggregation module. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI), pp. 1–5 (2022)
10. Khader, F., et al.: Medical diffusion-denoising diffusion probabilistic models for 3D medical image generation. arXiv preprint arXiv:2211.03364 (2022)
11. Lu, W., et al.: An investigation of quantitative accuracy for deep learning based denoising in oncological PET. Phys. Med. Biol. **64**(16), 165019 (2019)
12. Lu, Z., Li, Z., Wang, J., Shen, D.: Two-stage self-supervised cycle-consistency network for reconstruction of thin-slice MR images. arXiv preprint arXiv:2106.15395 (2021)
13. Luo, Y., et al.: 3D transformer-GAN for high-quality PET reconstruction. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12906, pp. 276–285. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87231-1_27
14. Luo, Y., et al.: Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis. Med. Image Anal. **77**, 102335 (2022)
15. Noroozi, M., Favaro, P.: Unsupervised learning of visual representations by solving jigsaw puzzles. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9910, pp. 69–84. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46466-4_5
16. Oktay, O., et al.: Attention U-Net: learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
17. Onishi, Y., et al.: Anatomical-guided attention enhances unsupervised PET image denoising performance. Med. Image Anal. **74**, 102226 (2021)
18. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695 (2022)
19. Slovis, T.L.: The ALARA concept in pediatric CT: myth or reality? Radiology **223**(1), 5–6 (2002)
20. Song, T., Yang, F., Dutta, J.: Noise2Void: unsupervised denoising of PET images. Phys. Med. Biol. **66**(21), 214002 (2021)
21. Wang, Y., et al.: 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. IEEE Trans. Med. Imaging **38**(6), 1328–1339 (2019)
22. Yan, J., Lim, J., Townsend, D.: MRI-guided brain PET image filtering and partial volume correction. Phys. Med. Biol. **60**(3), 961 (2015)
23. Yie, S., Kang, S., Hwang, D., Lee, J.: Self-supervised PET denoising. Nucl. Med. Mol. Imaging **54**(6), 299–304 (2020)
24. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
25. Zhang, X., et al.: Total-body dynamic reconstruction and parametric imaging on the uEXPLORER. J. Nucl. Med. **61**(2), 285–291 (2020)