*AI Medicine*
https://www.sciltp.com/journals/aim

*Article*

# Multi-Modality ViT-cGAN for Low-Dose PET Image Reconstruction

Yaling Fang [1], Jiahui Yang [1], Haoran Wan [2], Shuhan Jin [2], Shaoya Wang [2]
and Yueyang Teng [2,3,*]

[1] School of Electrical and Control Engineering, Shenyang Jianzhu University, Shenyang 110168, China
[2] College of Medicine and Biological Information Engineering, Northeastern University, Shenyang 110169, China
[3] Key Laboratory of Intelligent Computing in Medical Imag, Ministry of Education, Shenyang 110169, China
* Correspondence: tengyy@bmie.neu.edu.cn

**How To Cite**: Fang, Y.; Yang, J.; Wan, H.; et al. Multi-Modality ViT-cGAN for Low-Dose PET Image Reconstruction. *AI Medicine* **2026**, *3*(1), 1. https://doi.org/10.53941/aim.2026.100001

Received: 9 December 2025
Revised: 18 December 2025
Accepted: 29 December 2025
Published: 23 January 2026

**Abstract:** Positron emission tomography (PET) is an important medical imaging technique that reflects the molecular activity of tissues and organs by injecting radioactive tracers. Low-dose (LD) PET is gradually being adopted to reduce radiation dose and scanning costs, however this usually leads to increased image noise and artifacts, which can affect clinical diagnosis. Therefore, in order to maintain high-quality PET image generation while utilizing LD-PET data, this paper proposes a multi-modality Vision Transformer-based conditional generative adversarial network (ViT-cGAN) that directly achieves high-quality PET image reconstruction using the corresponding LD-PET sinogram data and computed tomography (CT) images. Specifically, the network incorporates the advantages of Vision Transformer and multi-modality inputs. In addition, an extensive objective function is designed to optimize the network for improving the details and visual quality of the reconstructed images. Experimental results show that our proposed method can effectively reconstruct high-quality PET images, outperforming current state-of-the-art methods.

**Keywords:** condition generative adversarial network (cGAN); image reconstruction; positron emission tomography (PET)

## 1. Introduction

Positron emission tomography (PET) is an important medical diagnosis technique for functional, metabolic, and molecular imaging of organs or tissues in modern nuclear medicine. It can reflect the activities of different organs and tissues at the molecular level through the injection of specific radiotracers into the living body [1,2]. However, PET scanning is expensive and this process also causes potential radiation damage to the human body. Low-dose (LD) sampling methods are often used, which can effectively reduce the radiation dose and also reduce the cost of scanning to some extent [3]. However, this can also lead to a reduction in the number of detected photons, creating noise and artifacts in the reconstructed image, which can negatively affect clinical diagnosis. Until now, a variety of conventional PET reconstruction methods have been proposed, including analytical reconstruction [4–7], iterative reconstruction [8–10], and post-reconstruction of the image domain [11–13]. They share common drawbacks: sensitivity to noise and poor image quality under LD data. Therefore, how to design algorithms to improve the quality of PET image reconstruction under LD sampling conditions is a problem that must be solved today.

In the past decade, deep learning has widely penetrated into several fields of medical imaging [14–17] and has been successfully applied to image reconstruction methods. Gong et al. [18] proposed an iterative reconstruction based on U-Net model to reconstruct PET images. Häggström et al. [19] proposed a deep encoder-decoder network that directly solves the PET image reconstruction problem. Spuhler et al. [20] proposed the reconstruction of full-count PET from low-count images using a dilated convolutional neural network.

In recent years, methods based on generative adversarial network (GAN) have further contributed to the development of medical image reconstruction, which generates high-quality images by simulating adversarial training with generators and discriminators, and are now also gradually used in PET reconstruction tasks [21–24]. Conditional GAN (cGAN) is an extended version of GAN, in which the core is to introduce conditional variables to guide the generation process. Since the generated images are more targeted, it can reduce the ambiguity and randomness between the generated samples, and make up for the shortcomings of ordinary GAN in the quality of the generated data and the stability of the training. It performs better than the ordinary GAN in many practical applications. Wang et al. [25] used cGAN for the first time to realize high quality PET image reconstruction. Liu et al. [26] proposed to use cGAN to reconstruct PET images directly from sinograms.

Despite the good potential of the cGAN based image reconstruction task, it still faces the following limitations in terms of generation quality and optimization efficiency. First, both generators and discriminators under the adversarial network framework usually process image data through convolutional operations, which makes it limited in modeling remote semantic dependencies in the data. Lacking non-local contextual information, reconstructed PET images may lose or obtain inaccurate global structure. Inspired by recent significant advances in the attentional mechanism for medical image analysis, visual transformer (ViT) [27] can directly capture feature interactions on a global scale [28,29]. In LD-PET images, ViT better restores global coherence and reduces the impact of noise on overall image quality. Moreover, PET image reconstruction is affected by the attenuation effect, i.e., the rays are absorbed or scattered as they pass through the body's tissues. Having information on the density of different types of tissue is crucial for correcting for attenuation [30,31], for example, CT image information can provide assistance with anatomical structures. However, most of the existing methods are unimodal, and relying only on a single source of data often fails to help in PET attenuation correction.

Consequently, in order to achieve high-quality PET reconstruction under LD sampling conditions, we improved the classical cGAN framework-based network Pix2Pix [32] and proposed a multi-modality perceptual ViT-cGAN network (MPeVitcGAN). It is able to reconstruct the high-quality PET image directly from the LD-PET sinogram data and the corresponding CT images. Specifically, we incorporate the ViT encoder structure in the generator to capture global semantic information and enhance the focus on the focal region. Next, we improve the generator to a multimodal input method with additional extraction of CT information, aiming to provide more comprehensive information and reduce the attenuation effect during PET reconstruction. In addition, we design an extensive objective function that combines the original adversarial loss, $L_1$ loss, and perceptual loss to optimize the network for generating more natural and realistic medical images. These methods not only improve the prediction ability of the cGAN model, but also more accurately guide the generator to produce more realistic PET images in critical regions and improves the visual quality of the generated images. The experimental results show that our method obtains the best reconstruction results compared with other state-of-the-art methods, in both quantitative and qualitative analyses, and the generated images possess better visual quality.

## 2. Methology

The overall framework of our proposed MPeVitcGAN is shown in Figure 1, which consists of two parts: generator network and discriminator network. First, the input data undergoes preprocessing via the Domain Transform (DT) layer. This module employs a back-projection operation to reconstruct simulated LD sinogram image into an intermediate domain transform image. The objective is to explicitly reduce discrepancies with the target image domain, thereby providing a more favourable starting point for subsequent generative networks. After that, the real PET image is used as the learning target of the DT map and the corresponding CT image. The PET image is generated by learning the mapping relationship between them through a modified generator. Specifically, the modified generator is a combination of the original U-Net-based generator of Pix2Pix with the architecture of the ViT encoder and the addition of a downsampling channel to extract the features of the CT image, which is called the ViT-UNet generator. We then use an adversarial learning strategy for the designed network, with the discriminator being Pix2Pix's original PatchGAN. This model is a discriminator based on local image blocks, which enhances the quality of generated details by evaluating local regions rather than the entire image. The network structure of the algorithm is described in detail below.

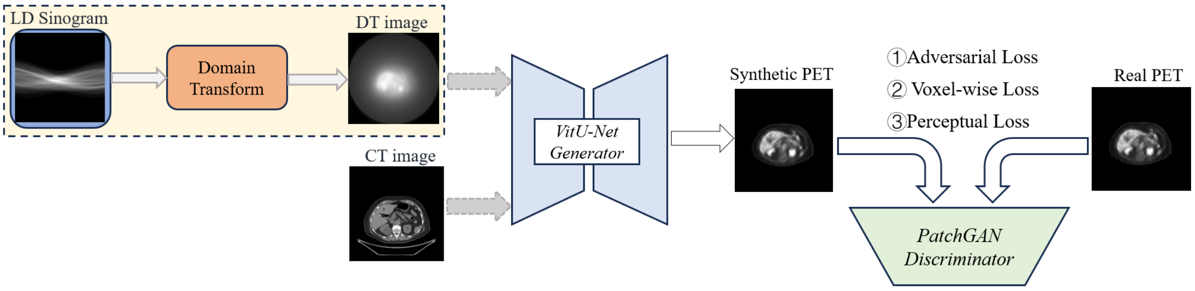

**Figure 1.** Proposed network architecture of MPeVitcGAN for LD sinogram PET data reconstruction.

### 2.1. ViT-UNet Generator Network

The generator architecture is crucial for the quality of synthesized images. Our proposed generator architecture, shown in Figure 2, consists of two sub-level networks: ViT encoder network (ViTEncoder) and Two-channel U-Net network (TCU-Net).

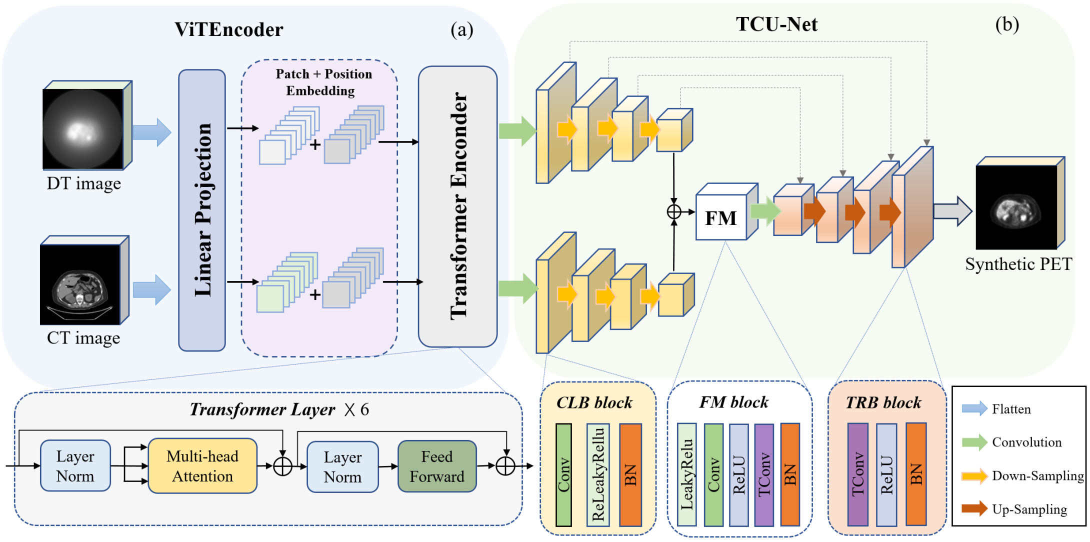

**Figure 2.** Network structure diagram of the ViT-UNet generator proposed by the MPeVitcGAN network directorate. Part (**a**) illustrates the ViTEncoder structure, while part (**b**) depicts the TCU-Net structure. FM—Feature Merging, CLB—Convolutional Layer Block, TRB—Transposed Convolutional Layer Block, Conv—Convolution, NB—Batch Normalization, ReLU—Rectified Linear Unit, TConv—Transposed Convolution.

ViTEncoder is used to convert the input images into sequence information and learn the remote dependencies between them for modeling. Figure 2a shows the detailed structure of ViTEndoer. The DT and the CT images are processed by the same ViTEncoder module, respectively, which first divides the input image into fixed-size patches, with the input image size of 128 × 128 and patch size of 16 × 16. Each image generates 64 (image size/patch size) patches, and each patch is linearly projected and then positional coding is added to obtain a set of one-dimensional sequences. These sequences are inputted into the Transformer encoder, which consists of six Transformer layers, and each layer contains respectively a multi-attention (MA) module and a feed-forward network (FFN) accompanied by layer normalization (LN) layer.

TCU-Net is used to extract local information from features, fuse different modalities, and restore the fused features to the final target image. One of its core tasks is to constrain the spatial distribution of PET functional data using CT anatomical information. During feature fusion, the distinct anatomical structural features extracted from the CT channel serve as spatial priors. They guide and modulate the feature responses of the PET channel to ensure that the reconstructed metabolic activity distribution remains confined within anatomically plausible regions.In addition, jump-joins are used for multi-level feature aggregation. As shown in Figure 2b, it consists of a two-channel downsampling module, a Feature Merging (FM) module and a single-channel upsampling module. The DT and CT image feature sequences processed by the ViTEncoder are first passed through a convolutional layer to adjust the number of channels, making them ready for subsequent downsampling. The convolutional kernel size used is 4 × 4 with a step size of 2. The two downsampling channels share a similar structure but differ in detailed parameters. The downsampled portion of each channel consists of four identical blocks, including the convolutional layer, the LeakyRelu activation function layer, and the batch normalization (BN) layer, called the CLB block. The output channels are 64, 128, 256, 512, respectively. The downsampled feature maps are subsequently merged and fed into the FM module. This module employs a multi-scale fusion strategy combining weighted summation with direct concatenation. Its core component is the learnable linear weighted fusion of same-scale features, expressed as follows:

$$\mathbf{F}_{\text{fuse}} = \sum_{i=1}^{N} w_i \cdot \mathbf{F}_i \tag{1}$$

where $w_i$ denotes the trainable weight, and $F_i$ represents the $i$-th input feature at the same scale. For cross-scale features, aggregation is achieved through channel concatenation. The fused features undergo deep integration and transformation via a convolutional layer, a transposed convolutional layer, ReLU and LeakyReLU activation layer and BN layer. The output channels are 1024. Finally, the image resolution is gradually restored by four layers of up-sampling network in which each layer includes transposed convolutional layer, Relu activation function and BN layer. The output channel is 512, 256, 128, 64, respectively. Finally, in the output layer, the normalization layer is removed, and the synthesized image is obtained by the transposed convolutional layer with 7 × 7 size and the tanh  function.

### *2.2. PatchGAN Discriminator Network*

The PatchGAN discriminator is a key component of the Pix2Pix network used to determine the authenticity of a generated image. Unlike traditional discriminators, PatchGAN does not classify the image as a whole, instead, it splits the image into multiple small patches, each of which is individually determined to be a real image or not. This approach not only allows the discriminator to focus on the details of localized regions of the image, but also effectively improves the accuracy of training and helps the generator to reconstruct the details more accurately.

The specific structure is shown in Figure 3, using a typical convolutional neural network (CNN) architecture. There are four convolutional layers in which each layer includes a convolutional layer, LeakyReLU activation function and LN layer. The filter size of $4 \times 4$, padding of 1 and step 2. The number of kernels in the four convolutional layers are 64, 128, 256, 512, respectively. After these four convolutional layers, the last convolutional operation uses the $4 \times 4$ convolutional kernel with a step size of 1. The output of this convolutional layer is a single-channel feature map that is used to predict whether each patch belongs to the real image. This means that each position on each output feature map corresponds to a patch of an image and gives a binary classification result as to whether the patch is a real image or not. In this way, the PatchGAN is able to evaluate the quality of the generated images at a finer granularity, especially the ability to capture local details. It makes the generated images more realistic in terms of local structure and texture, and thus promoting better training of the generator.

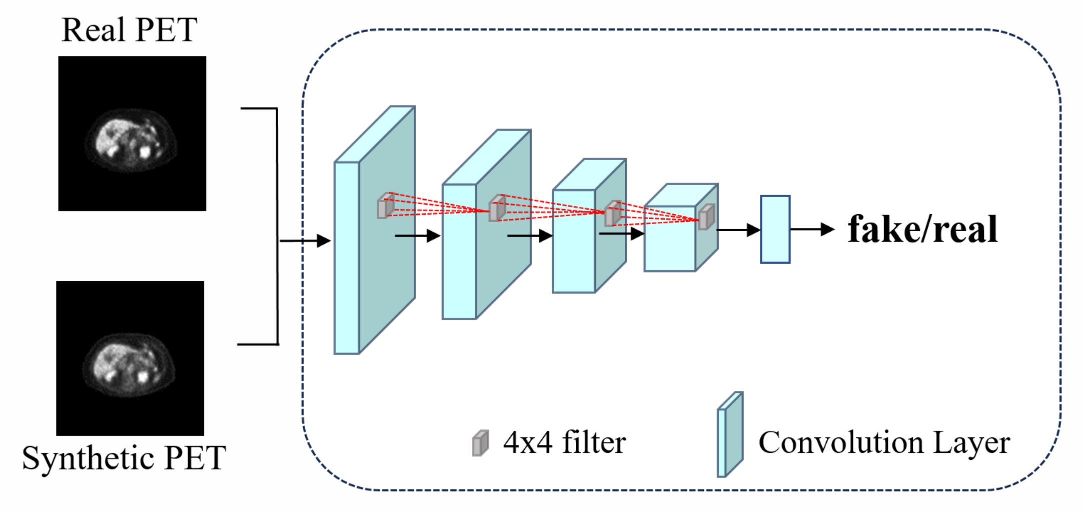

**Figure 3.** The discriminator network proposed in this method.

### *2.3. Loss*

The success of image synthesis requires semantic reasoning, and for the task of PET image reconstruction, the synthesized PET outputs must be semantically similar to the corresponding inputs, albeit with dramatic changes in appearance. To further improve the quality of the generated PET images, an extensive objective functions are used to optimize the network, including three types of terms: adversarial loss, pixel-level reconstruction loss, and perceptual loss.

The generator of a traditional GAN generates images without additional input conditions, and its main goal is to produce images that can deceive the discriminator, which is responsible for distinguishing between real and generated images. Although this approach can produce diverse generated images, it has weak control over the generated results and it is difficult to ensure the accuracy of the images under specific conditions. Conditional GAN guides the generation process of the generator and the discriminator by taking additional information (e.g., labels, modalities, or noise) as input conditions. This conditional information enables the generator to produce images that better meet specific requirements, while the discriminator takes these conditions into account when evaluating image veracity. Given a DT image $x_{DT} \in R_{DT}$, CT image $x_{CT} \in R_{CT}$, and the corresponding PET image $y \in R_{PET}$, reconstructed PET image G ($x_{DT}$, $x_{CT}$) is generated from $x_{DT}$ and $x_{CT}$ by generator network. The conditional adversarial loss can be defined as:

$$\begin{aligned}\mathcal{L}_{adv}(D) =& \frac{1}{2}E[(D(x_{DT}, y) - 1)^2] \\ &+ \frac{1}{2}E[(D(x_{DT}, G(x_{DT}, x_{CT}))))^2]\end{aligned} \tag{2}$$

$$\mathcal{L}_{adv}(G) = \frac{1}{2}E[(D(x_{DT}, G(x_{DT}, x_{CT}))) - 1)^2] \tag{3}$$

To ensure that the reconstructed PET images are close to their corresponding real images, we employ $L_1$ loss as a voxel-by-voxel estimation error to narrow the gap between them, which means that the generator not only needs to deceive the discriminator, but also needs to minimize the absolute pixel intensity difference between the synthesized PET images and the real PET images. The $L_1$ pixel-level reconstruction loss is formalized as follows:

$$\mathcal{L}_{L1}(G) = E\left\| y - G(x_{DT}, x_{CT}) \right\|_1 \tag{4}$$

Although pixel-level loss captures the overall structure, it does not reflect the perceptual difference between the synthesized image and the real image. As an example, consider two identical PET images which, if they differ by only one pixel, would be significantly different in terms of pixel-level loss, despite being perceptually similar [33]. Therefore, the introduction of perceptual loss based on the similarity of high-level feature representations can generate higher quality PET images, which is formalized as follows:

$$\mathcal{L}_{per}(G) = E\left\| V(y) - V(G(x_{DT}, x_{CT})) \right\|_1 \tag{5}$$

where *V* is the set of feature mappings before the second maxpooling operation of the pretrained VGG-16 [34].

Our overall MPeVitcGAN objective function is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{adv}(G) + \mathcal{L}_{adv}(D) + \alpha\mathcal{L}_{per}(G) + \beta\mathcal{L}_1(G) \tag{6}$$

The parameters $\alpha$ and $\beta$ are balances between different losses, and we empirically set $\alpha$ = 1 and $\beta$= 1.

## 3. Experiments

We use a radiogenomic dataset built from a non-small cell lung cancer (NSCLC) cohort consisting of 211 subjects [35], with paired CT and PET images for each patient, and scanned 32,786 2D slices. The LD sinogram data is generated by projecting these 2D slices using a systematic matrix orthogonal projection, and then transformed into visualized DT images by an inverse Radon transformation to invert the orthoprojection process and transform the projected data into visualized DT images. In order to reduce the computational cost, we resized the PET and CT images to 128 × 128. In order to fully utilize the available data, the dataset was divided into ten, of which seven were used as the training set, one was used as the validation set, and the remaining two were used as the test set.

To evaluate the performance of the proposed MPeVitcGAN, we use three evaluation metrics to validate the reconstruction results. The first metric is the Peak Signal-to-Noise Ratio (PSNR) [36], where higher values indicate better image quality.

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{\max^2}{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}\right) \tag{7}$$

We use structural similarity (SSIM) as a second metric, which is used for visual image quality assessment and considers the overall structure of the image. The value of this metric ranges from 0 to 1, with higher values indicating that the image structure is closer to the real image:

$$\mathrm{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

where $\mu_x$ and $\mu_y$ are the means of *x* and *y*, $\sigma_x^2$ and $\sigma_y^2$ are the variances of *x* and *y* respectively, $\sigma_{xy}$ is the covariance between *x* and *y*, $C_1$ and $C_2$ are the small constants avoiding zero denominator errors.

The third evaluation metric is the relative root mean square error (rRMSE), where lower values indicate better image quality.

$$\mathrm{rRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}}{\bar{y}} \tag{9}$$

Our proposed MPeVitcGAN is implemented by PyTorch, and the model is trained and tested on an Intel (R) Core (TM) i7-10700 2.90 GHz GPU with 8 GB of memory. Depending on the computer hardware used, the training batch size was set to 8 and the learning rate to 0.00002.

### *3.1. Hyperparameter Selection*

Figure 4 shows the convergence of the loss curve on the training set during the training process of the MPeVitcGAN. When the network training process reaches the 200-th epoch, the loss curve on the validation set no longer decreases, and we stop the training in order to avoid network overfitting.

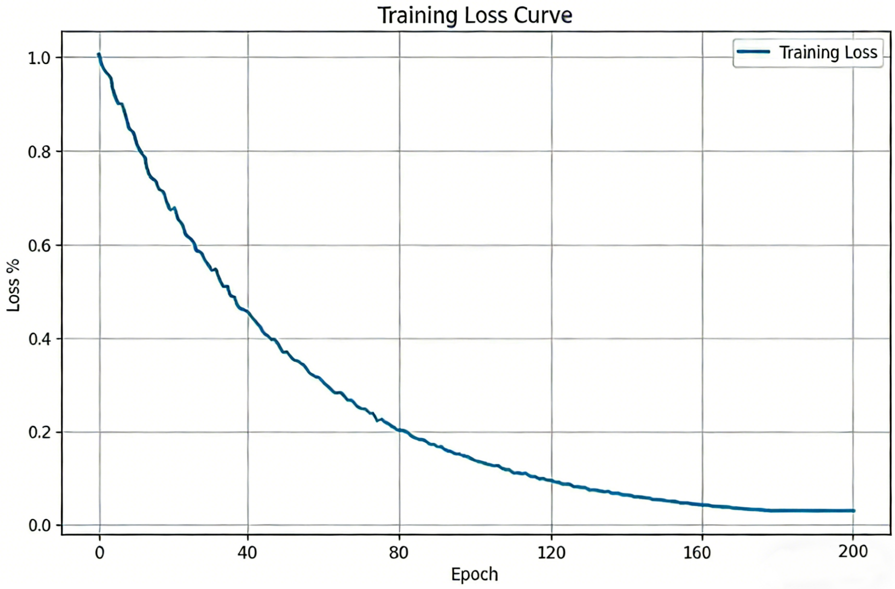

**Figure 4.** The loss curve between the reconstructed image and the real image.

In order to reduce the complexity of the proposed model while maintaining the optimal model performance, we first investigated the effect of the number of transformer layers on the performance of the final model and estimated the optimal layer settings. The statistical results expressed in terms of PSNR, SSIM and rRMSE are shown in Figure 5. It can be seen that the proposed model achieves the best results in PSNR and rRMSE when the number of transformer layers is 6, although its SSIM is slightly lower than that of the 7-layer configuration. Considering that PSNR and rRMSE are more critical metrics for evaluating pixel-level fidelity in medical image reconstruction, the 6-layer design is deemed superior. Furthermore, owing to the approximately linear relationship between the parameter count and the number of layers in Transformer architectures, the 6-layer model reduces theoretical computational cost by about 14.3% compared to the 7-layer model, leading to higher deployment efficiency. By balancing core performance and computational efficiency, the 6-layer architecture represents the optimal trade-off and is selected as our final design.

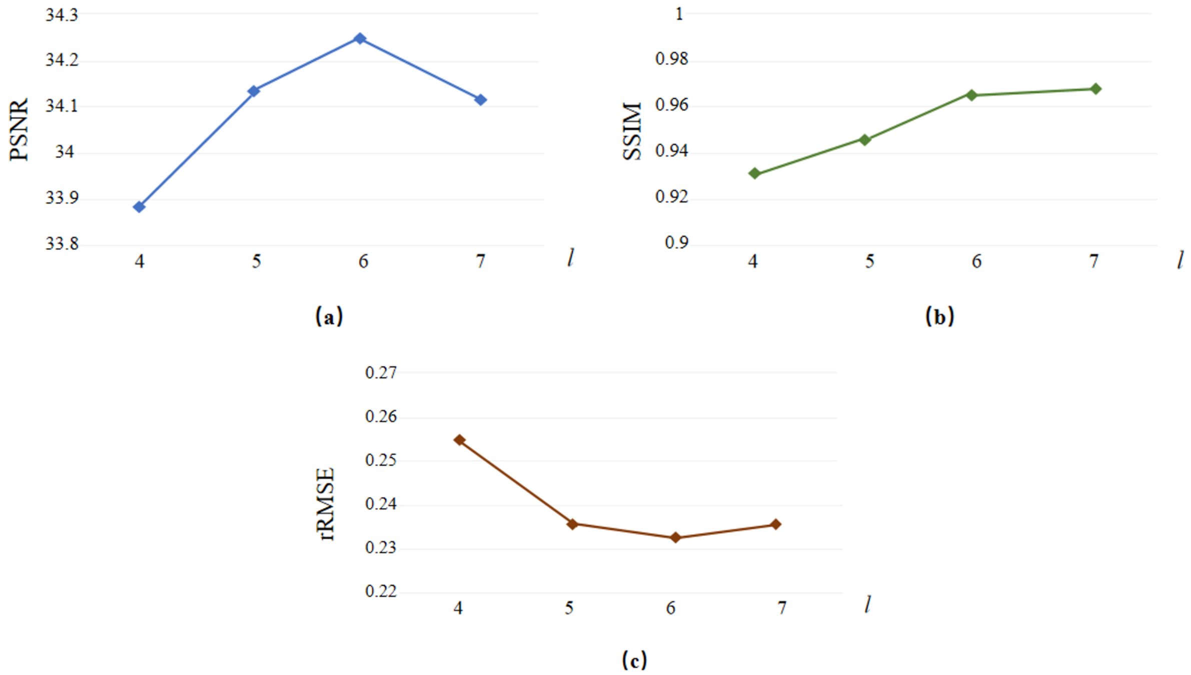

**Figure 5.** Comparison of (**a**) PSNR (**b**) SSIM and (**c**) rRMSE for different number of transformer layers.

### *3.2. Ablation Stuies*

Our enhancement of the original Pix2Pix network can be divided into two parts: the ViT-UNet generator and the generalized loss function. To study the contribution of each part, we designed and performed ablation experiments.

#### 3.2.1. Contribution of the ViT-UNet Generator

As described earlier, we designed the ViT-UNet generator to optimize the network, and in order to evaluate the effectiveness of this generator, we trained each of these improvements individually against the structure. The network that used only U-Net as the original generator is denoted as baseline. The network that added ViTEncoder to U-Net for feature extraction of DT images is denoted as VitcGAN. The network that used TCU-Net for multimodal inputs to both DT and CT images is denoted as McGAN. And a comparison is made with our method. The specific PSNR, SSIM and rRMSE comparison results are shown in Table 1.

**Table 1.** Quantitative comparison of our proposed MPeVitcGAN model with respect to its three variants.

| **Method** | **PSNR** | **SSIM** | **rRMSE** |
|---|---|---|---|
| Baseline | 33.147 | 0.954 | 0.284 |
| VitcGAN | 33.463 | 0.956 | 0.277 |
| McGAN | 34.014 | 0.962 | 0.245 |
| Ours | 34.229 | 0.965 | 0.237 |

In order to validate the benefits of the ViTEncoder, we compare the baseline with the VitcGAN. The only difference between these two models is whether the DT image is first processed by the VitEncoder. As can be seen in the first and second rows of Table 1, the results of the three evaluation metrics are optimized when the ViTEncoder is included in the baseline. Among them, PSNR (SSIM) improved from 33.147 (0.954) to 33.463 (0.956), and rRMSE decreased from 0.284 to 0.277. From the second and third columns of Visual Comparison in Figure 6, the distribution of metabolic activities within the high-intensity region is more accurately in the generated PET images due to the ability of ViT's global attention mechanism to focus on important regions while preserving global coherence. The above results indicate that ViTEncoder can further improve the image quality of synthesized PET images.

To verify the benefits of the TCU-Net module in providing multimodal inputs, we compared baseline with McGAN. The first and third rows of Table 1 give the quantitative comparison results. It can be seen that the method improves on all three performance metrics. Specifically, there is an improvement of 0.867 PSNR, 0.008 SSIM, and a decrease of 0.039 in rRMSE. Observing the second and fourth columns of Figure 6, the reconstructed PET images have clearer anatomical boundaries, higher spatial resolution, and are more similar to the real images due to the high-resolution anatomical structure information provided by the CT images. These results demonstrate the effectiveness of the multimodal approach.

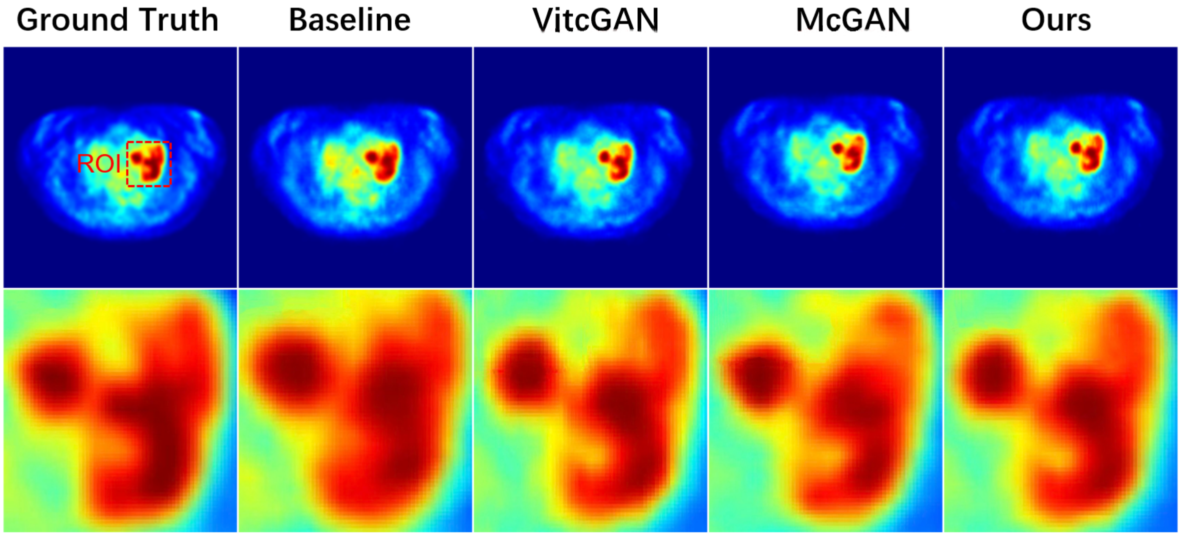

**Figure 6.** Ablation experiments in Grund Truth and images generated using different methods, with the second row showing the method region of interest (ROI) for the red boxed region marked in the first row.

3.2.2. Contribution of the Extensive Loss Function

As described in the Methods section, our proposed an extensive objective function to compute the loss includes the adversarial loss, the $L_1$ loss, and the perceptual loss. In order to investigate its effectiveness for PET reconstruction task, we make the above losses separately to train the proposed model. The results of quantitative comparison of PSNR, SSIM and rRMSE between different loss functions are shown in Table 2.

**Table 2.** Quantitative comparison of PSNR, SSIM and rRMSE between different loss functions.

| Loss Function | PSNR | SSIM | rRMSE |
|---|---|---|---|
| Adversarial | 32.859 | 0.945 | 0.306 |
| Adversarial + $L_1$ | 33.975 | 0.958 | 0.246 |
| Ours | 34.229 | 0.965 | 0.237 |

We note that the quality of the reconstructed images is not as good in the model with only adversarial loss. The main problem lies in the fact that the adversarial loss usually does not guarantee that the generated image is exactly the same as the target image at the pixel level, and the use of only the adversarial loss may lead to a loss of image details. The first extension is to add the $L_1$ loss, from the second row of the table, we can see that all the three indexes have increased, which proves that this loss can effectively make up for the shortcomings of the adversarial loss that can not guarantee the consistency at the pixel level. PSNR and rRMSE are pixel-level metrics, which align with the optimization objective of the $L_1$ loss, making it effective in improving these two metrics. However, $L_1$ loss has limited ability to recover image structure and texture, so the improvement effect on SSIM may be small. The second extension is to add the perceptual loss function, which just makes up for the gap of the network to improve the visual perceptual quality of images, so the improvement of SSIM index is significant.

### *3.3. Performance Comparison*

In order to evaluate the effectiveness and superiority of our network, we compare the proposed MPeVitcGAN with four state-of-the-art methods including Pix2Pix, LCPR-Net, SAGAN and ResViT. Specifically Pix2Pix is a preliminary version of this work, LCPR-Net also reconstructs the PET images directly from the corresponding sinogram data, SAGAN is a convolution-based GAN model who improves by incorporating a self-attention module in the generator. And ResViT explores the combination of Residual Networks and ViT for medical image synthesis. For a fair comparison, the number of attention modules in SAGAN and ResViT is the same as in this paper.

The average quantitative comparison results for the subjects with NSCLC are shown in Table 3, respectively. It can be seen that our method obtains the best PSNR, SSIM, and rRMSE in general, which can effectively improve the quality of PET reconstructed images. Specifically, in subjects with NSCLC, MPeVitcGAN outperforms the second-best method by 1.082 in PSNR (≈3.26%) and 0.007 in SSIM (≈0.7%). More critically, rRMSE is reduced from 0.284 to 0.237, which corresponds to a 16.5% reduction. These concurrent improvements validate the superior reconstruction quality of our proposed network.

**Table 3.** Quantitative comparison of PSNR, SSIM and rRMSE between different reconstruction methods.

| Method | PSNR | SSIM | rRMSE |
|---|---|---|---|
| SAGAN | 31.644 | 0.952 | 0.349 |
| ResViT | 32.706 | 0.957 | 0.318 |
| LCPR-Net | 32.574 | 0.958 | 0.317 |
| Pix2Pix | 33.147 | 0.954 | 0.284 |
| MPeVitcGAN (Ours) | 34.229 | 0.965 | 0.237 |

In addition to the quantitative comparison results, we also provide the visual comparison results of the first slice in Figure 7. The first row is a gray-scale map generated by Ground Truth as well as five different methods, the second row is a transformed pseudo-color map, and the third row is the corresponding error map. Compared with the other compared methods, the reconstructed PET images generated by the methods in this paper are richer in details and have sharper edges, especially in the regions pointed by the boxes. It can also be found from the corresponding error maps that the error maps of the proposed MPeVitcGAN are darker in color, which means that the PET images generated by our proposed network have the smallest difference from the ground truth. In addition to that, Figure 8 shows the intensity distribution (profile) maps of pixel values in vertical and horizontal directions of the first slice, respectively. It can be seen that the method proposed in this paper is the closest to the true value, which indicates that the PET images reconstructed by the network are more in line with the metabolic distribution of the actual lesions and are more accurate in recovering the intensity of metabolic activities and anatomical features. As shown in Figures 9 and 10, we also present the visual comparison and profile plots of the second slice, where our method also demonstrates excellent results. Overall, compared with other state-of-the-art methods, our method possesses better reconstruction results, and the generated images are able to better maintain the edge contours and provide high-quality visualization.

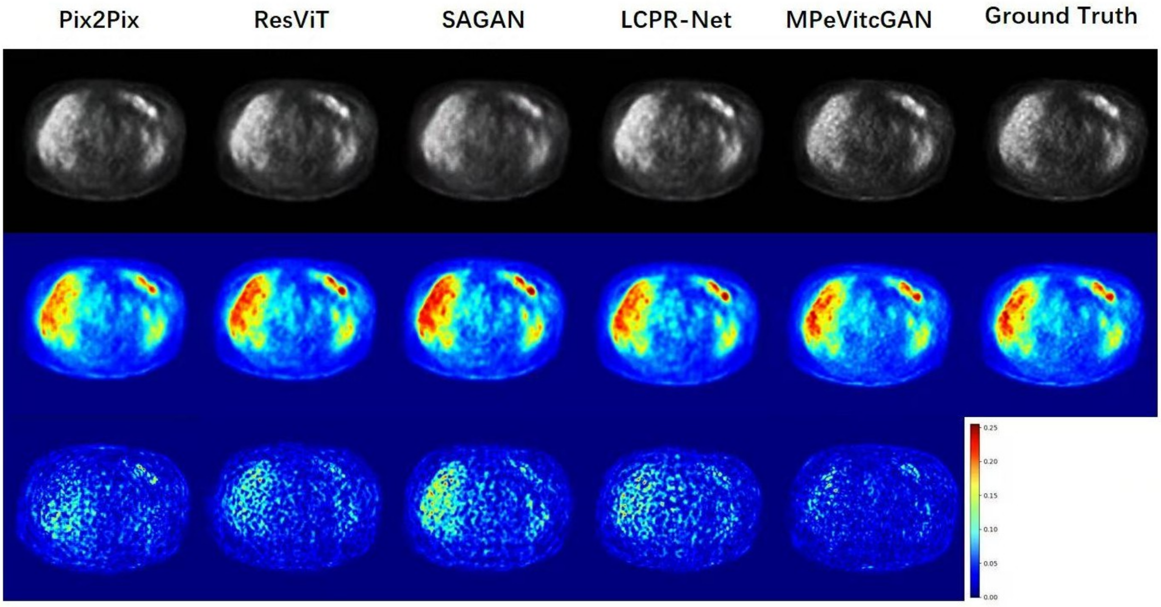

**Figure 7.** The visual comparison image of Slice 1 for different methods of PET reconstruction. The first row shows Ground Truth as well as the grayscale maps generated by the five different methods, the second row shows the transformed pseudo-color maps, and the third row shows the corresponding error maps.

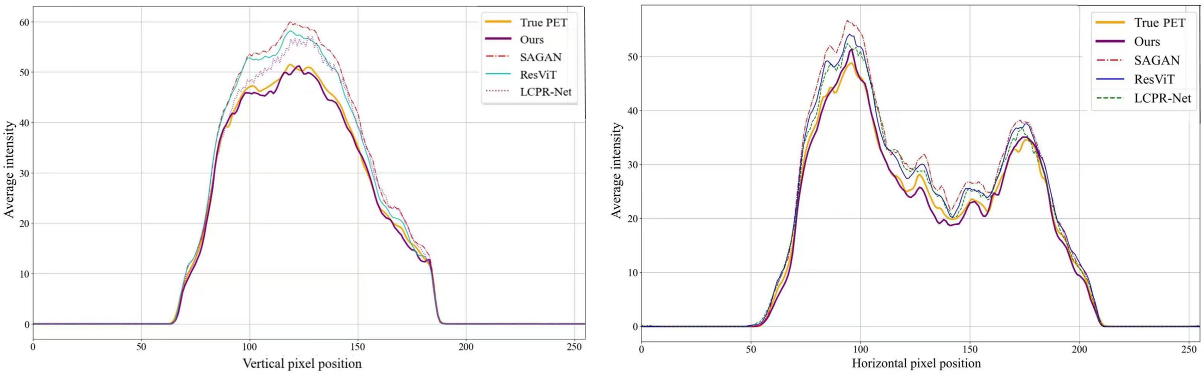

**Figure 8.** The profile plots of Slice 1 for different methods and the ground truth in vertical and horizontal directions.

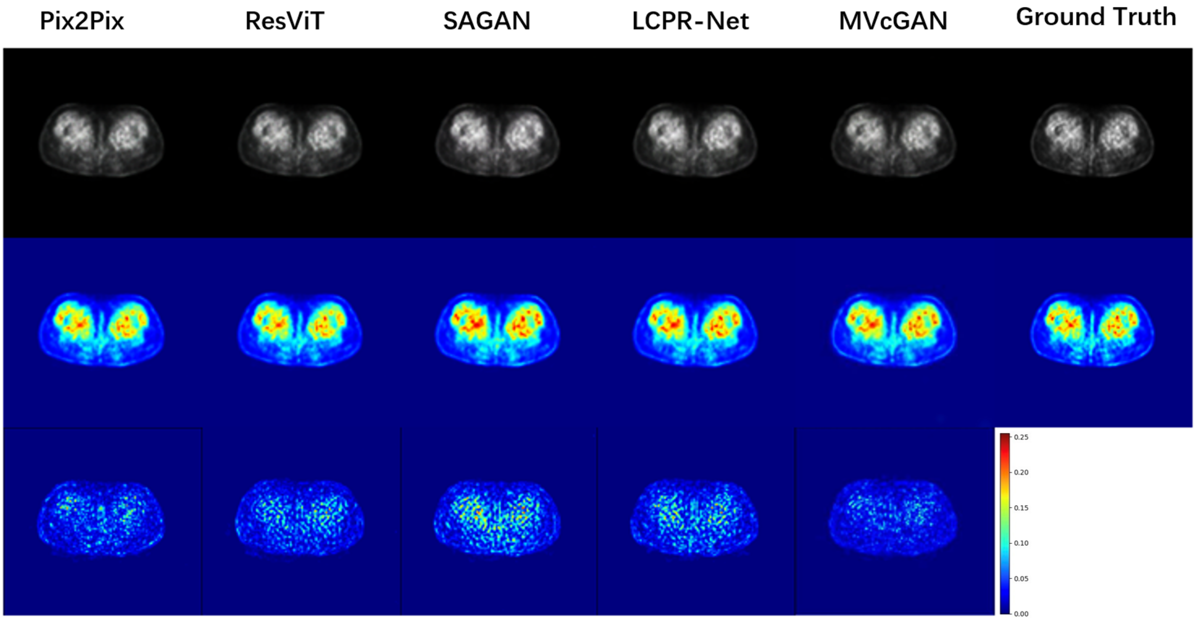

**Figure 9.** The visual comparison image of Slice 1 for different methods of PET reconstruction.

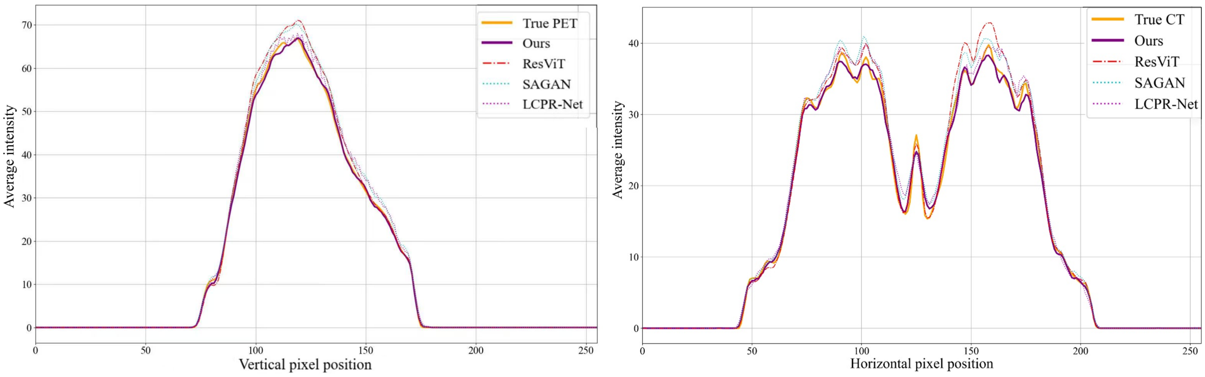

**Figure 10.** The profile plots of Slice 2 for different methods and the ground truth in vertical and horizontal directions.

## 4. Conclusions

In this paper, we proposed a multimodal perceptual cGAN aimed at reconstructing high-quality PET images from LD simulated PET sinusoidal images and corresponding CT images. The method can provide physicians with high-quality PET images and reduce the injection dose of radiotracer and scanning time. The experimental results show the significant superiority of our proposed method compared to other state-of-the-art methods.

Although the proposed MPeVitcGAN has made significant progress in LD-PET image reconstruction, there are still some limitations. The training process of the network requires a large multi-modal dataset, and the acquisition and annotation of such data can be limited, particularly in clinical settings. Future research can address these issues in the following directions. On one hand, more efficient data augmentation methods can be explored to reduce the reliance on large annotated datasets. On the other hand, the method can be extended to other types of medical image reconstruction tasks, such as the fusion and reconstruction of functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) signals, to further validate its broad applicability in the multi-modal imaging field. In summary, while the proposed method demonstrates significant advantages, there remains considerable room for future research and development.

## Author Contributions

Y.F.: software, validation, writing—reviewing and editing; J.Y.: data curation; H.W.: conceptualization, methodology; S.J.: visualization; S.W.: investigation; Y.T.: supervision; All authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Institutional Review Board Statement

Ethical review and approval were waived for this study because it performed a secondary analysis of the publicly available, de-identified NSCLC Radiogenomics dataset. The original data collection and sharing were conducted in accordance with ethical standards, as documented on The Cancer Imaging Archive (TCIA) project page: https://www.ncbi.nlm.nih.gov/pubmed/30325352 (accessed on: 18 December 2025). Our analysis of this anonymized data posed no additional risk to the subjects.

## Informed Consent Statement

This study analyzed the publicly available NSCLC Radiogenomics dataset. As the data are fully anonymized and publicly accessible for research purposes, the requirement for informed consent were waived for this secondary analysis.

## Data Availability Statement

The original datasets analyzed in this study are publicly available in the NSCLC Radiogenomics collection of The Cancer Imaging Archive (TCIA) at the persistent URL: https://www.cancerimagingarchive.net/collection/nsclc-radiomics/ (accessed on: 18 December 2025 ).

## Conflicts of Interest

The authors declare no conflict of interest.

## Use of AI and AI-Assisted Technologies

No AI tools were utilized for this paper.

## References

1. Chen, W. Clinical applications of PET in brain tumors. *J. Nucl. Med.* **2007**, *48*, 1468–1481.
2. Wang, Y.; Zhang, P.; An, L.; et al. Predicting standard-dose PET image from low-dose PET and multimodal MR images using mapping-based sparse representation. *Phys. Med. Biol.* **2016**, *61*, 791.
3. Wang, Y.; Ma, G.; An, L.; et al. Semisupervised tripled dictionary learning for standard-dose PET image prediction using low-dose PET and multimodal MRI. *IEEE Trans. Biomed. Eng.* **2016**, *64*, 569–579.
4. Tang, X.; Ning, R. A cone beam filtered backprojection (CB-FBP) reconstruction algorithm for a circle-plus-two-arc orbit. *Med. Phys.* **2001**, *28*, 1042–1055.
5. Tao, X.; Zhang, H.; Wang, Y.; et al. VVBP-tensor in the FBP algorithm: Its properties and application in low-dose CT reconstruction. *IEEE Trans. Med. Imaging* **2019**, *39*, 764–776.
6. Balda, M.; Hornegger, J.; Heismann, B. Ray contribution masks for structure adaptive sinogram filtering. *IEEE Trans. Med. Imaging* **2012**, *31*, 1228–1239.
7. Manduca, A.; Yu, L.; Trzasko, J.D.; et al. Projection space denoising with bilateral filtering and CT noise modeling for dose reduction in CT. *Med. Phys.* **2009**, *36*, 4911–4919.
8. Dutta, J.; Leahy, R.M.; Li, Q. Non-Local Means Denoising of Dynamic PET Images. *PLoS ONE* **2013**, *8*, e81390
9. Cong, Y.; Zhang, S.; Lian, Y. K-SVD Dictionary Learning and Image Reconstruction Based on Variance of Image Patches.

In Proceedings of the 2015 8th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 12–13 December 2015.

10. Yu, X.; Wang, C.; Hu, H.; et al. Low Dose PET Image Reconstruction with Total Variation Using Alternating Direction Method. *PLoS ONE* **2016**, *11*, e0166871.
11. Buades, A.; Coll, B.; Morel, J.M. A Non-Local Algorithm for Image Denoising. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005.
12. Zhang, H.; Ma, J.; Wang, J.; et al. Statistical image reconstruction for lowdose CT using nonlocal means-based regularization. Part II: An adaptive approach. *Comput. Med. Imaging Graph.* **2015**, *43*, 26–35.
13. Fumene Feruglio, P.; Vinegoni, C.; Gros, J.; et al. Block matching 3D random noise filtering for absorption optical projection tomography. *Phys. MedBiol.* **2010**, *55*, 5401–5415.
14. Moeskops, P.; Viergever, M.A.; Mendrik, A.M.; et al. Automatic segmentation of MR brain images with a convolutional neural network. *IEEE Trans. Med. Imaging* **2016**, *35*, 1252–1261.
15. Chen, H.; Zhang, Y.; Zhang, W.; et al. Low-dose CT via convolutional neural network. *Biomed. Opt. Express* **2017**, *8*, 679–694.
16. Gong, K.; Guan, J.; Liu, C.C.; et al. PET image denoising using a deep neural network through fine tuning. *IEEE Trans. Radiat. Plasma Med. Sci.* **2018**, *3*, 153–161.
17. Lei, Y.; Harms, J.; Wang, T.; et al. MRI-only based synthetic CT generation using dense cycle consistent generative adversarial networks. *Med. Phys.* **2019**, *46*, 3565–3581.
18. Gong, K.; Guan, J.; Kyungsang, K.; et al. Iterative PET Image Reconstruction Using Convolutional Neural Network Representation. *IEEE Trans. Med. Imaging* **2019**, *38*, 675–685.
19. Häggström, I.; Schmidtlein, C.R.; Campanella, G.; et al. DeepPET: A deep encoder-decoder network for directly solving the PET image reconstruction inverse problem. *Med. Image Anal.* **2019**, *54*, 253–262.
20. Spuhler, K.; Serrano-Sosa, M.; Cattell, R.; et al. Full-count PET recovery from low-count image using a dilated convolutional neural network. *Med. Phys.* **2020**, *47*, 4928–4938.
21. Wang, Y.; Zhou, L.; Wang, L.; et al. Locality Adaptive Multi-Modality GANs for High-Quality PET Image Synthesis. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, 16–20 September 2018.
22. Xue, H.; Zhang, Q.; Zou, S.; et al. LCPR-Net: Low-count PET image reconstruction using the domain transform and cycle-consistent generative adversarial networks. *Quant. Imaging Med. Surg.* **2021**, *11*, 749.
23. Luo, Y.; Zhou, L.; Zhan, B.; et al. Adaptive rectification based adversarial network with spectrum constraint for high-quality PET image synthesis. *Med. Image Anal.* **2022**, *77*, 102335 .
24. Fei, Y.; Zu, C.; Jiao, Z.; et al. Classification-aided high-quality PET image synthesis via bidirectional contrastive GAN with shared information maximization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022.
25. Wang, Y.; Yu, B.; Wang, L.; et al. 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. *Neuroimage* **2018**, *174*, 550–562.
26. Liu, Z.; Ye, H.; Liu, H. Deep-learning-based framework for PET image reconstruction from sinogram domain. *Appl. Sci.* **2022**, *12*, 8118.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arxiv* **2020**, arxiv:2010.11929.
28. Zhang, Z.; Yu, L.; Liang, X.; et al. TransCT: Dual-path transformer for low dose computed tomography. In Proceedings of the Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, 27 September–1 October 2021.
29. Zheng, H.; Lin, Z.; Zhou, Q.; et al. Multi-transsp: Multimodal transformer for survival prediction of nasopharyngeal carcinoma patients. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer Nature: Cham, Switzerland, 2022.
30. De Wever, W.; Ceyssens, S.; Mortelmans, L.; et al. Additional value of PET-CT in the staging of lung cancer: Comparison with CT alone, PET alone and visual correlation of PET and CT. *Eur. Radiol.* **2007**, *17*, 23–32.
31. Fletcher, J.W.; Kymes, S.M.; Gould, M.; et al. A comparison of the diagnostic accuracy of 18F-FDG PET and CT in the characterization of solitary pulmonary nodules. *J. Nucl. Med.* **2008**, *49*, 179–185.
32. Isola, P.; Zhu, J.Y.; Zhou, T.; et al. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
33. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016.
34. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Bakr, S.; Gevaert, O.; Echegaray, S.; et al. A radiogenomic dataset of non-small cell lung cancer. *Sci. Data* **2018**, *5*, 180202. https://doi.org/10.1038/sdata.2018.202.
36. Hore, A.; Ziou, D. Image Quality Metrics: PSNR vs. SSIM. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010.