Spach Transformer: Spatial and Channel-wise Transformer Based on Local and Global Self-attentions for PET Image Denoising

Se-In Jang, Tinsu Pan, Ye Li, Pedram Heidari, Junyu Chen, Quanzheng Li, and Kuang Gong

Abstract—**Position emission tomography (PET) is widely used in clinics and research due to its quantitative merits and high sensitivity, but suffers from low signal-to-noise ratio (SNR). Recently convolutional neural networks (CNNs) have been widely used to improve PET image quality. Though successful and efficient in local feature extraction, CNN cannot capture long-range dependencies well due to its limited receptive field. Global multi-head self-attention (MSA) is a popular approach to capture long-range information. However, the calculation of global MSA for 3D images has high computational costs. In this work, we proposed an efficient spatial and channel-wise encoder-decoder transformer, Spach Transformer, that can leverage spatial and channel information based on local and global MSAs. Experiments based on datasets of different PET tracers, i.e., ^{18}F-FDG, ^{18}F-ACBC, ^{18}F-DCFPyL, and ^{68}Ga-DOTATATE, were conducted to evaluate the proposed framework. Quantitative results show that the proposed Spach Transformer can achieve better performance than other reference methods.**

Index Terms—**Positron Emission Tomography, Image Denoising, Transformer, Self-attention**

I. Introduction

POSITRON emission tomography (PET) is an imaging modality widely used in oncology, neurology, and cardiology studies [1]–[6]. It can observe molecular-level activities *in vivo* through the injection of specifically designed radioactive tracers. Due to various physical degradation factors and limited counts received, the signal-to-noise ratio (SNR) of PET is inferior to other imaging modalities, which comprises its clinical values in diagnosis, prognosis, staging and treatment monitoring. Further enhancing PET image quality is essential to improve its diagnosis and treatment-monitoring accuracy.

For the past decades, various post-processing methods have been investigated to further improve PET image quality. Gaussian filtering is widely used in clinical scanners to reduce the noise but it can also smooth out image structures. Non-local mean [7], wavelet [8], highly constrained back-projection processing [9], [10], and guided filtering [11] have been proposed to further preserve image details based on pre-defined filters or similarity calculation. Recently, the convolutional neural network (CNN)-based approaches have become a dominant trend in various PET research topics such as image denoising [12]–[21] and image reconstruction [22]–[27], due to their better performance than other state-of-the-art methods. The CNN-based Unet architecture [28] is the most widely used network architecture currently. Performance of CNN for PET image denoising can be further improved by different loss-function designs, e.g., a perceptual loss [13], [29], [30] formed by a pre-trained network or a learned discriminative loss from the additional discriminator network [29]–[36].

One pitfall of CNN is that it specifically focuses on local spatial information and thus has a limited receptive field. The multi-head self-attention (MSA)-based structure, i.e., transformer, has the ability to capture long-range information [37]–[39]. Based on the aggregation of tokens, global MSAs were first developed for natural language processing [37]. Vision Transformer (ViT) was then proposed where global MSAs were successfully applied to sequences of spatial image patches for image classification [38]. Due to global MSA's quadratically growing computational complexity along the spatial dimension, Swin Transformer [40] was proposed to efficiently calculate local MSAs using shifted windows that offered linear computational complexity. Swin Transformer calculated local MSAs on multiple stages and achieved better classification performance than global MSAs. However, local MSAs of Swin Transformer is still limited regarding the receptive field of MSA. Restormer [41] was recently proposed to efficiently compute global MSAs along the channel dimension with linear computational complexity for image restoration tasks. Regarding medical imaging, various transformer-based architectures were investigated for medical image segmentation [42]–[47]. For PET image denoising, transformer-based architectures have not yet been fully investigated.

In this work, we proposed a *spa*tial and *ch*annel-wise encoder-decoder *transformer*, denoted as *Spach Transformer*, that can leverage spatial and channel information from oncological PET images. The proposed Spach Transformer includes spatial-wise and channel-wise transformer blocks, which can efficiently perform image denoising while also preserving image details, e.g., small lesion uptakes. The spatial-wise block is based on local attention and the channel-wise block

This work was supported by the National Institutes of Health under grants R03EB030280, R21AG067422, R01AG078250 and P41EB022544. Corresponding author: Kuang Gong (email: kgong@mgh.harvard.edu).

S. Jang, Q. Li and K. Gong are with Gordon Center for Medical Imaging and Center for Advanced Medical Computing and Analysis , Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

T. Pan is with Department of Imaging Physics, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Y. Li is with Center for Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

P. Heidari is with Division of Nuclear Medicine and Molecular Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

J. Chen is with Department of Electrical and Computer Engineering and Department of Radiology, Johns Hopkins University, Baltimore, MD, USA

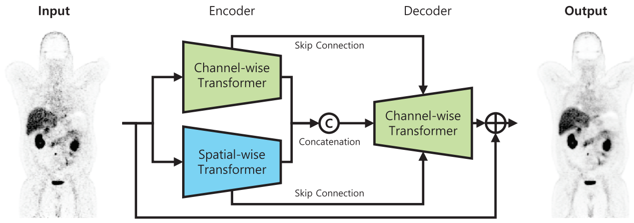


Fig. 1: A brief overview of Spach Transformer for PET image denoising.

is based on global attention. As the channel-wise transformer block may ignore the spatial information, a gated-conv feed-forward network (GCFN) block was further designed to incorporate spatial attributes between layers in the channel-wise transformer block. A brief overview of the proposed Spach Transformer is presented in Fig. 1. It encoded the input and built two deep latent features in spatial and channel directions. The decoder further took the two latent features concatenated as input to generate the final denoised PET image based on skip connections. To evaluate the effectiveness of the proposed Spach Transformer for PET image denoising, clinical ^{18}F-FDG, ^{18}F-ACBC, ^{18}F-DCFPyL, and ^{68}Ga-DOTATATE datasets were utilized to quantitatively compare the performance of different CNN and transformer-based methods.

The main contributions of this work include : (1) proposing a novel spatial and channel-wise encoder-decoder transformer integrating local and global attentions for PET image denoising and tumor-uptake preserving; (2) designing a GCFN block to incorporate spatial attributes in the channel-wise transformer block to boost the performance; (3) evaluations using clinical datasets based on different PET tracers. This paper is organized as follows. Section 2 presents the related works and the proposed Spach Transformer in detail. Experiments, results and discussions are shown in Sections 3, 4 and 5, respectively. Finally, conclusions are drawn in Section 5.

II. Method

Fig. 2 shows the detailed diagram of the proposed Spach Transformer. The encoder part of Spach Transformer consists of two levels of spatial-wise transformer blocks and four levels of channel-wise transformer blocks. The decoder part of Spach Transformer includes four levels of channel-wise transformer blocks. Below we will describe major elements of the Spach Transformer (Sec. II-A) as well as details of the spatial-wise transformer block (Sec. II-B) and the channel-wise transformer block. (Sec. II-C).

A. *Spach Transformer: Spatial and Channel-wise Transformer*

Given a low-dose PET image $\mathbf{X} \in \mathbb{R}^{H \times W \times D \times 1}$, Spach Transformer first obtained low-level feature embeddings $\mathbf{F}_0 \in \mathbb{R}^{H \times W \times D \times C}$ based on a $3 \times 3 \times 3$ convolution, where H, W, and D indicate the voxel dimension and C is the number of channels. The four layers of the channel-wise encoder transformer blocks further processed the embedding $\mathbf{F}_0$ to extract deep latent features $\mathbf{F}_{l_c}$.

Addtionally, following [38], [40], a patch partitioning module produced non-overlapping patches $\mathbf{x}_i \in \mathbb{R}^{P^3}$ from the low-dose PET image $\mathbf{X}$, where P is the patch size and was set to 4 in our implementation. A linear embedding layer was added to project these non-overlapping voxels to an arbitrary dimension as follows:

$$\mathbf{z}_0 = [\mathbf{x}_0\mathbf{E}; \mathbf{x}_1\mathbf{E}; \dots ; \mathbf{x}_{N_p}\mathbf{E}] \in \mathbb{R}^{N_P \times d}, \tag{1}$$

where $\mathbf{E} \in \mathbb{R}^{P^3 \times d}$ is a projection matrix, and $N_P = \frac{H}{P} \times \frac{W}{P} \times \frac{D}{P}$ is the number of patches. The spatial-wise encoder transformer blocks and the patch-merging layer used the tokens $\mathbf{z}_0$ to extract deep latent features $\mathbf{F}_{l_s}$. Role of the patch-merging layer is to reduce the spatial size by half along each dimension.

The encoder part described above gradually reduced the spatial size while expanding the channel capacity. The final latent features were generated by concatenating the latent features from the two transformer paths as $\mathbf{F}_l = [\mathbf{F}_{l_c}, \mathbf{F}_{l_s}]$. The final latent features $\mathbf{F}_l$ were further supplied to the decoder path to generate the high-quality PET image $\hat{\mathbf{X}}$. Following [41], a $1 \times 1 \times 1$ convolution and a concatenation operation was utilized to reduce the channel size by half at each decoder layer, followed by a channel-wise transformer block to recover the spatial information. The deep features $\mathbf{F}_d$ were further refined to obtain robust features $\mathbf{F}_r$ at the PET image resolution scale. Finally, a $3 \times 3 \times 3$ convolution layer was applied on $\mathbf{F}_r$ to produce the residual image $\mathbf{R}$, and the final denoised PET image was generated as $\hat{\mathbf{X}} = \mathbf{R} + \mathbf{X}$.

B. *Spatial-wise Transformer Block*

Swin Transformer [40] proposed to use window-based and shift window-based MSA modules to replace the standard MSA modules. Fig. 3a shows the diagram of the transformer block, which is based on the following operations:

$$\begin{aligned}
\hat{\mathbf{z}}_l &= MSA_W\left(LN(\mathbf{z}_{l-1})\right) + \mathbf{z}_{l-1}, \\
\mathbf{z}_l &= MLP\left(LN(\hat{\mathbf{z}}_l)\right) + \hat{\mathbf{z}}_l, \\
\hat{\mathbf{z}}_{l+1} &= MSA_{SW}\left(LN(\mathbf{z}_l)\right) + \mathbf{z}_l, \\
\mathbf{z}_{l+1} &= MLP\left(LN(\hat{\mathbf{z}}_{l+1})\right) + \hat{\mathbf{z}}_{l+1},
\end{aligned} \tag{2}$$

where MSA_W and MSA_{SW} are regular window-based MSA (W-MSA) and shifted window-based MSA (SW-MSA), respectively, MLP is the multi-layer perceptron (MLP) module with the Gaussian error linear unit (GELU) [48] as the activation layer, and LN is the layer normalization (LN) module [49] which was inserted before the W-MAS, SW-MSA and MLP modules. Residual connection was used after each module. Fig. 3b shows the window-based self-attention diagram. The self-attention was calculated as [40] :

$$\text{Attention}(\mathbf{Q}_s, \mathbf{K}_s, \mathbf{V}_s) = \text{Softmax}\left(\mathbf{Q}_s\mathbf{K}_s^T/\sqrt{d} + \mathbf{B}\right)\mathbf{V}_s, \tag{3}$$

where $\mathbf{B} \in \mathbb{R}^{M^3 \times M^3}$ is the relative position of tokens in each window, d is the query/key dimension. and M^3 is the number of patches in the 3D image window. The query $\mathbf{Q}_s$, key $\mathbf{K}_s$, and value $\mathbf{V}_s$ matrices were computed as:

$$\mathbf{Q}_s = W_l^Q\mathbf{Y}, \mathbf{K}_s = W_l^K\mathbf{Y}, \mathbf{V}_s = W_l^V\mathbf{Y}, \tag{4}$$

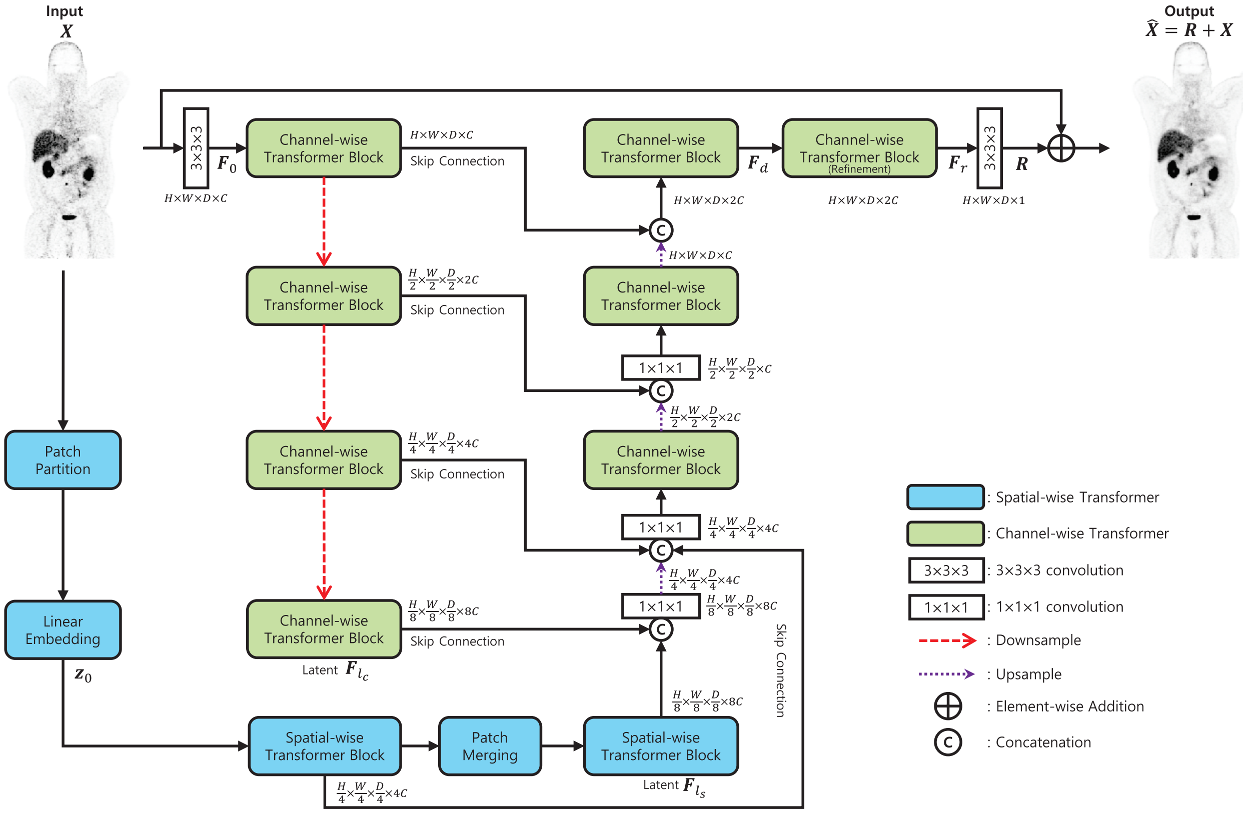


Fig. 2: Diagram of the proposed Spach Transformer for PET image denoising. Details of the channel-wise transformer block and the spatial-wise transformer block are given in Fig. 3 and Fig. 4, respectively.

where $W_l^Q \in \mathbb{R}^{M^3\times M^3}$, $W_l^K \in \mathbb{R}^{M^3\times M^3}$, and $W_l^V \in \mathbb{R}^{M^3\times M^3}$ are the linear projection matrices, and $\mathbf{Y} \in \mathbb{R}^{M^3\times d}$ is the input after the LN layer.

C. Channel-wise Transformer Block

The Restormer work [41] designed the multi-Dconv head transposed attention (MDTA) modules to replace the standard MSA modules, which computed self-attention along the channel dimension instead of the spatial dimension. This can allow a global attention calculation with linearly computational complexity. In Restormer, each transformer block included a MDTA and a gated-Dconv feed-forward network (GDFN) module. Though Restormer showed competing performance in natural image enhancement, based on our experiments, we found that this configuration has the pitfall of losing local spatial information. Apart from transferring local spatial information in parallel using the spatial-wise transformer blocks as described in Sec. II-B, in the proposed channel-wise transformer block (see Fig. 4a), we further designed a gated-conv feed-forward network (GCFN) to better capture local spatial information and thus improve the lesion-uptake recovery. Below we will introduce the details of MDTA, GDFN and GCFN.

1) MDTA: Fig. 4b shows the diagram of the MDTA module. Given an input feature map $\mathbf{F} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}\times\hat{C}}$, a layer normalization module was first applied to obtain $\mathbf{Y} \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}\times\hat{C}}$. The query $\mathbf{Q}_c \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}\times\hat{C}}$, key $\mathbf{K}_c \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}\times\hat{C}}$, and value $\mathbf{V}_c \in \mathbb{R}^{\hat{H}\times\hat{W}\times\hat{D}\times\hat{C}}$ matrices were then obtained after a $1\times1\times1$ pixel-wise convolution operation (encoding channel-wise context) and a $3\times3\times3$ channel-wise convolution operation (aggregating pixel-wise cross-channel context). By reshaping operations, $\hat{\mathbf{Q}}_c \in \mathbb{R}^{\hat{H}\hat{W}\hat{D}\times\hat{C}}$, $\hat{\mathbf{K}}_c \in \mathbb{R}^{\hat{C}\times\hat{H}\hat{W}\hat{D}}$, and $\hat{\mathbf{V}}_c \in \mathbb{R}^{\hat{H}\hat{W}\hat{D}\times\hat{C}}$ were obtained from $\mathbf{Q}_c$, $\mathbf{K}_c$, and $\mathbf{V}_c$, respectively, and were used to generate a transposed-attention map $\mathbf{A} \in \mathbb{R}^{\hat{C}\times\hat{C}}$ instead of the massive spatial attention map of size $\hat{H}\hat{W}\hat{D}\times\hat{H}\hat{W}\hat{D}$ [37], [38]. The output feature of the MDTA module $\hat{\mathbf{F}}$ can be expressed as

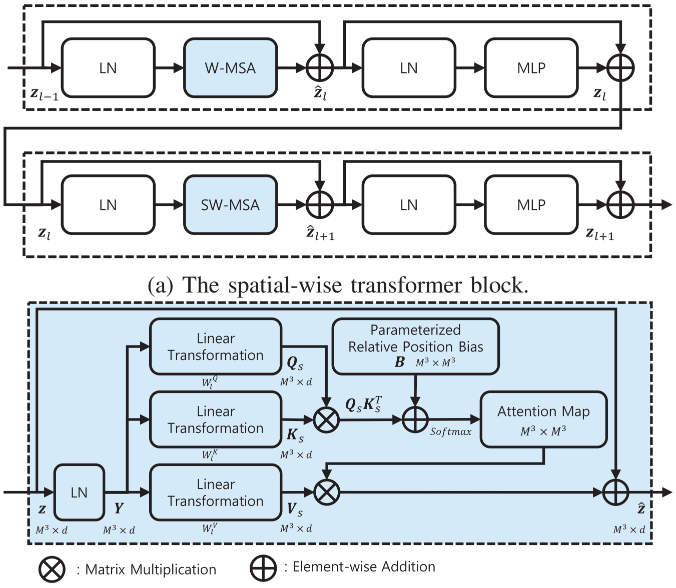
$$\begin{aligned} &\hat{\mathbf{F}} = W_p \text{Attention}(\hat{\mathbf{Q}}_c, \hat{\mathbf{K}}_c, \hat{\mathbf{V}}_c) + \mathbf{F}, \\ &\text{Attention}(\hat{\mathbf{Q}}_c, \hat{\mathbf{K}}_c, \hat{\mathbf{V}}_c) = \hat{\mathbf{V}}_c \cdot \text{Softmax}(\hat{\mathbf{Q}}_c \cdot \hat{\mathbf{K}}_c / \alpha), \end{aligned} \tag{5}$$

where α is a learnable scaling parameter to control the magnitude of the dot product output $\hat{\mathbf{Q}}_c \cdot \hat{\mathbf{K}}_c$, and W_p is a linear projection matrix. Note that the output of the self-attention operation was reshaped to have the same size as the input $\mathbf{F}$.

2) GDFN: A gating mechanism [50] can help selectively pass useful information to the next layer in the network architecture. Due to this, the gating network could suppress less informative features. In [41], the gating mechanism was utilized in the GDFN module (see Fig. 4c), which was calculated as

$$\begin{aligned} &\hat{\mathbf{F}} = W_p^0 \text{Gating}(\mathbf{F}) + \mathbf{F}, \\ &\text{Gating}(\mathbf{F}) = \phi\left(W_c^1 W_p^1 LN(\mathbf{F})\right) \odot W_c^2 W_p^2 LN(\mathbf{F}). \end{aligned} \tag{6}$$

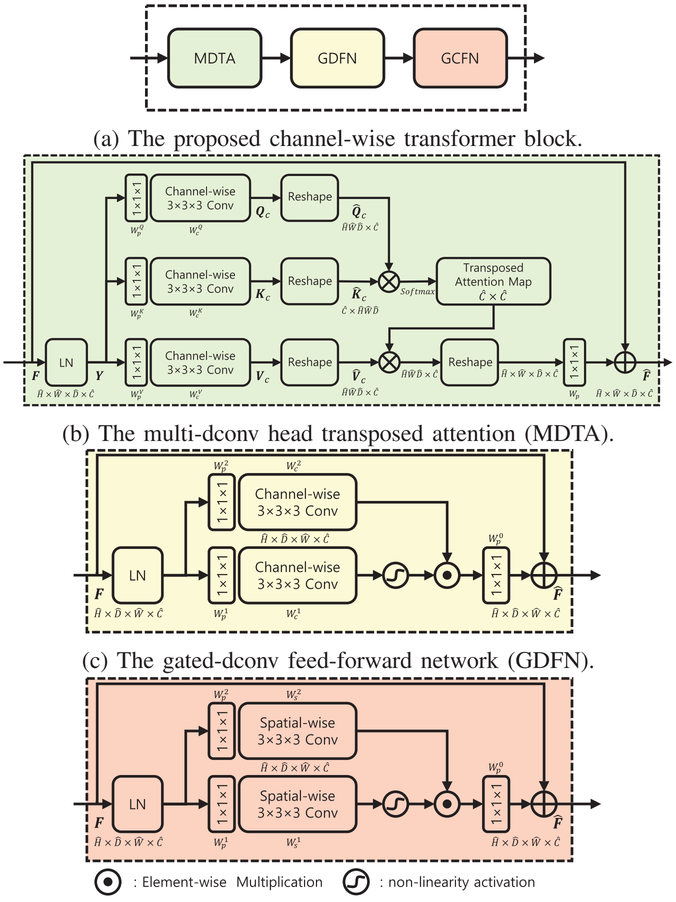
Here $\mathbf{F}$ and $\hat{\mathbf{F}}$ are the input and output features, $\odot$ is an element-wise multiplication, ϕ is the GELU non-linearity activation layer, and LN indicates the LN layer. W_p^0, W_p^1,



(a) The spatial-wise transformer block.

(b) The window-based multi-head self-attention.

Fig. 3: Diagrams of (a) the spatial-wise transformer block and (b) the window-based attention calculation.


MDTA
GDFN
GCFN
LN
F
$\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}$
Y
1×1×1
W_p^Q
Channel-wise 3×3×3 Conv
W_c^Q
Q_c
Reshape
$\hat{Q}_c$
$\hat{H}\hat{W}\hat{D} \times \hat{C}$
1×1×1
W_p^K
Channel-wise 3×3×3 Conv
W_c^K
K_c
Reshape
$\hat{K}_c$
$\hat{C} \times \hat{H}\hat{W}\hat{D}$
Softmax
Transposed Attention Map
$\hat{C} \times \hat{C}$
1×1×1
W_p^V
Channel-wise 3×3×3 Conv
W_c^V
V_c
Reshape
$\hat{V}_c$
$\hat{H}\hat{W}\hat{D} \times \hat{C}$
$\hat{H}\hat{W}\hat{D} \times \hat{C}$
Reshape
$\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}$
1×1×1
W_p
$\hat{F}$
$\hat{H} \times \hat{W} \times \hat{D} \times \hat{C}$
W_p^2
W_c^2
LN
F
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
1×1×1
Channel-wise 3×3×3 Conv
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
1×1×1
W_p^1
Channel-wise 3×3×3 Conv
W_c^1
W_p^0
1×1×1
$\hat{F}$
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
W_p^2
W_s^2
LN
F
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
1×1×1
Spatial-wise 3×3×3 Conv
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
1×1×1
W_p^1
Spatial-wise 3×3×3 Conv
W_s^1
W_p^0
1×1×1
$\hat{F}$
$\hat{H} \times \hat{D} \times \hat{W} \times \hat{C}$
: Element-wise Multiplication
: non-linearity activation

(a) The proposed channel-wise transformer block.

(b) The multi-dconv head transposed attention (MDTA).

(c) The gated-dconv feed-forward network (GDFN).

(d) The gated-conv feed-forward network (GCFN).

Fig. 4: Diagrams of (a) the proposed channel-wise transformer block, (b) the MDTA module, (c) the GDFN module and (d) the GCFN module.

TABLE I: According to channel number change C, an ablation study of different model parameters of Spach Transformer (GCFN).

	C	Model Parameters	PSNR		SSIM	
			Avg	Std	Avg	Std
Spach Transformer	6	7,034,951	54.9753	3.6359	0.9367	0.0147
	12	19,218,323	55.0840	3.6092	**0.9403**	**0.0128**
	24	59,350,043	55.0911	3.6103	0.9358	0.0147
	36	120,501,731	55.1003	3.6193	0.9379	0.0137
	48	202,673,387	**55.1443**	**3.6036**	0.9368	0.0138

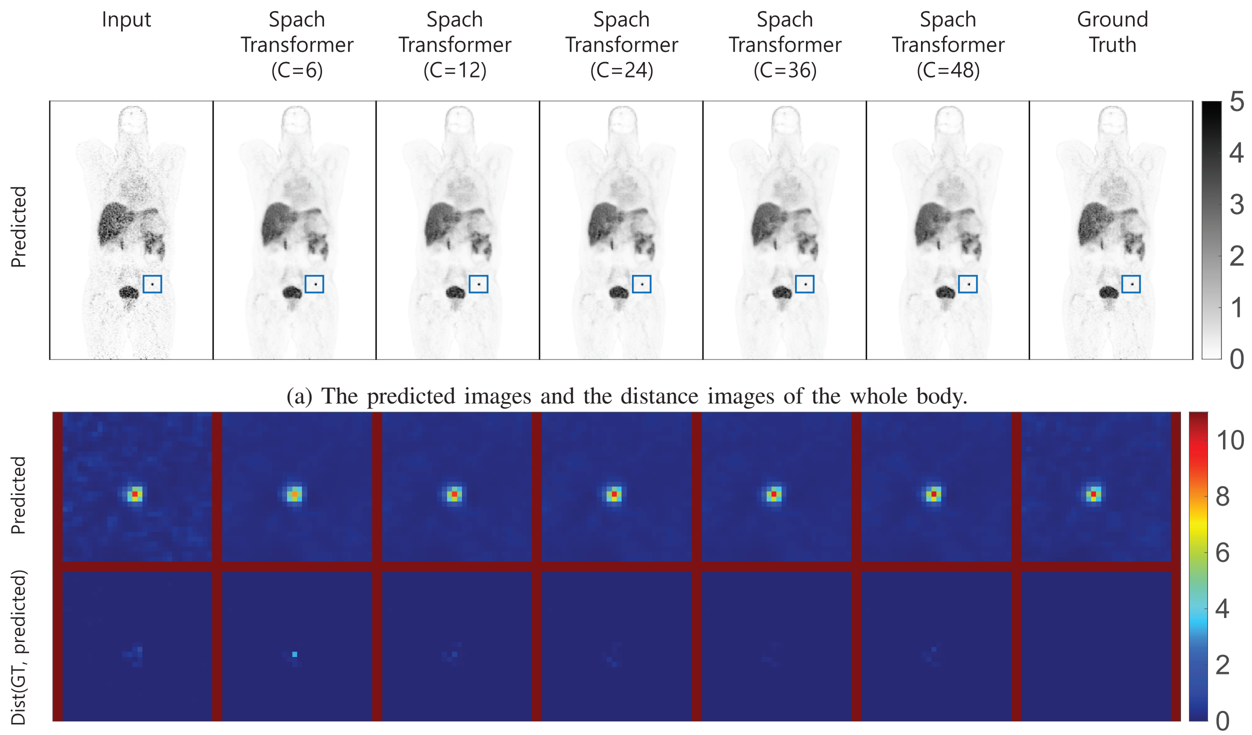
and W_p^2 indicate the linear projection matrices. W_c^1 and W_c^1 represent $3 \times 3 \times 3$ channel-wise convolutions.

3) GCFN: Since the above MDTA and GDFN modules entirely focus on utilizing channel information for image denoising, it may lose spatial information. To preserve spatial attributes between layers, a gated-conv feed-forward network (GCFN) module (see Fig. 4d) was added to the channel-wise transformer block. The only difference between GCFN and GDFN are that the $3 \times 3 \times 3$ channel-wise convolutions used in GDFN (W_c^1 and W_c^1) were replaced by $3 \times 3 \times 3$ spatial-wise convolutions (W_s^1 and W_s^1). We hypothesized that the spatial-wise convolutions could better utilize spatial information and thus further help recovery lesion uptakes.

III. Experiments

A. Datasets

To evaluate performance of the proposed Spach Transformer, clinical whole-body ^{18}F-FDG, ^{18}F-ACBC, ^{18}F-DCFPyL, and ^{68}Ga-DOTATATE datasets acquired from the GE DMI PET/CT scanner were utilized. These tracers have unique distributions in the organs and are currently the most widely used tracers in oncology PET studies. For each dataset, 1/4 of the events were extracted from the listmode data to generate the 1/4 low-dose data, which was reconstructed and supplied as the network input. The corresponding normal-dose image was adopted as the training label. Both the low-dose and normal-dose PET images were reconstructed based on the ordered subset expectation maximization (OSEM) algorithm (3 iterations and 17 subsets) with point spread function (PSF) and time of flight (TOF) modeling. For the training set, 30 ^{18}F-FDG and 30 ^{18}F-ACBC datasets were included. For the validation set, 4 ^{18}F-FDG datasets were used. For the testing set, 10 ^{18}F-FDG, 10 ^{18}F-ACBC, 10 ^{18}F-DCFPyL, and 18 ^{68}Ga-DOTATATE datasets were employed. Note that ^{18}F-DCFPyL and ^{68}Ga-DOTATATE datasets were not included in the training set, but in the testing set. This is to evaluate the robustness of the deep learning structures to datasets of new tracers unseen during the training phase. The whole-body PET images had a dimension of $256 \times 256 \times 345$ and a voxel size of $2.73 \times 2.73 \times 2.8$ mm^3. During network training, the input image was divided into $96 \times 96 \times 96$ cubes due to the GPU memory limit.



(a) The predicted images and the distance images of the whole body.

(b) The enlarged images where the blue rectangles are located in (a).

Fig. 5: Coronal and zoomed-in views of the denoised PET images of Spach Transformer according to different numbers of channels C based on an ^{18}F-DCFPyL dataset

B. Implementation and quantification

The Unet [28], Swin Transformer-based structure [47], and Restormer [41] were adopted as the reference methods. All the architectures utilized the Charbonnier loss [51] as the training loss function. For Spach Transformer, the channel size of the transformer blocks can influence the network parameters a lot. We have investigated the performance of the Spach Transformer with different channel sizes. During network training, the AdamW optimizer was used and 300 epochs were run with the initial learning rate of 1e-5 gradually reduced to 1e-8 using cosine annealing [52]. The proposed network was implemented on the Nvidia Tesla V100 GPU.

As for quantification, the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) were utilized to quantify the global performance of different methods. For oncological PET images, preserving tumor uptake is essential for disease diagnosis and monitoring. To evaluate the performance of different methods regarding tumor-uptake recovery, the contrast-to-noise (CNR) was calculated at the tumor region as

$$\text{CNR} = (V_{\text{tumor}} - V_{\text{ref}})/\sigma_{\text{ref}}. \tag{7}$$

Here V_{tumor} is the mean value of the tumor region of interest (ROI). V_{ref} and σ_{ref} are the reference ROIs' mean value and standard deviation, respectively. The reference ROIs were selected from uniform regions inside the liver region.

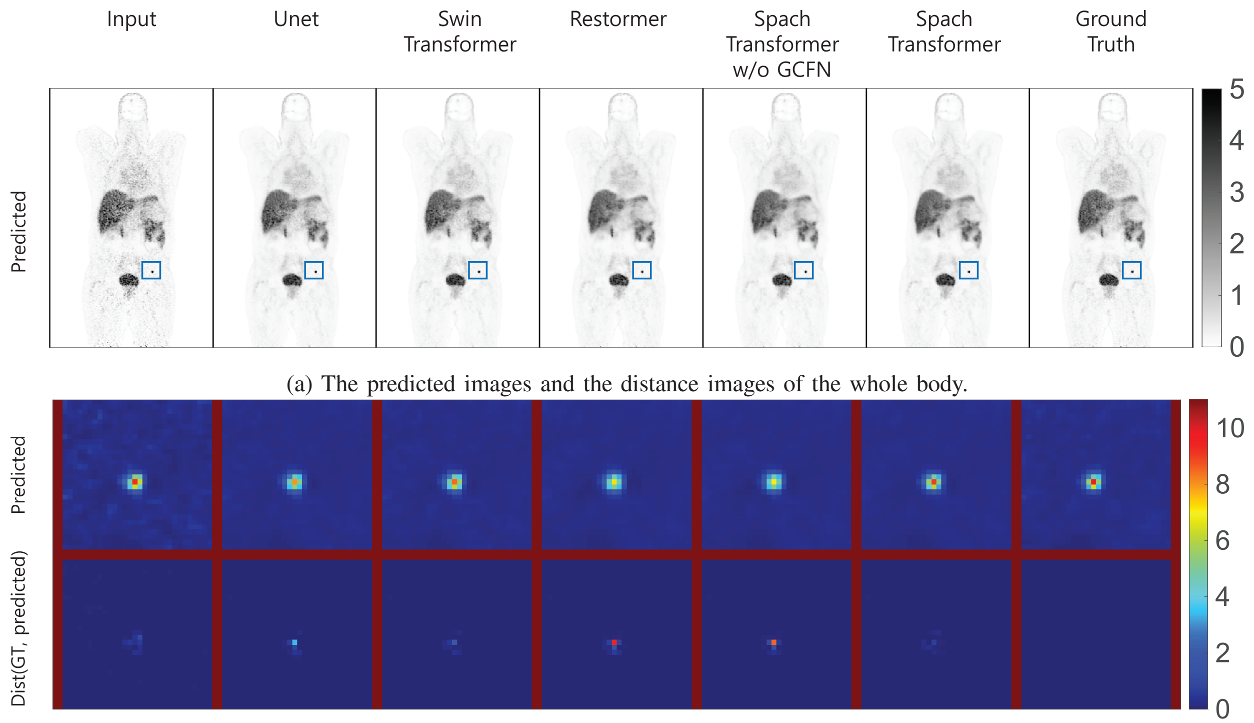
IV. Results

A. Effect of model parameters

We want to first investigate the performance of Spach Transformer with different network parameters. The channel number used in the transformer architecture has a big impact on the model size and was varied to investigate the network performance with different network parameters. Table I shows the averaged PSNR and SSIM values for different model parameters of the Spach Transformer based on varying the channel size C ($C = \{6, 12, 24, 36, 48\}$). It can be observed that the best PSNR was achieved when $C = 48$, while the best SSIM was obtained when $C = 12$. Fig. 5 shows the denoised PET images from one ^{18}F-DCFPyL dataset with different channel numbers. To evaluate how close the predicted images are to the ground truth (GT), the images of $L2$-norm distance between the predicted and GT images were also generated. It can be observed that when $C = 6$, the denoised image is blurred and the uptake in the zoomed-in tumor region cannot be fully recovered. When the channel number C was increased to 12, the network can better recover the tumor uptake and also has a better denoising performance. After further increasing the network parameter, performance of the network on tumor-uptake recovery increased a little and became stable. Based on the PSNR and SSIM results as well as the image appearance, the channel size C was set to 12 in all our studies. The comparison methods were all set to have a similar number of network parameters as the Spach Transformer with $C = 12$.

B. Comparisons with reference methods

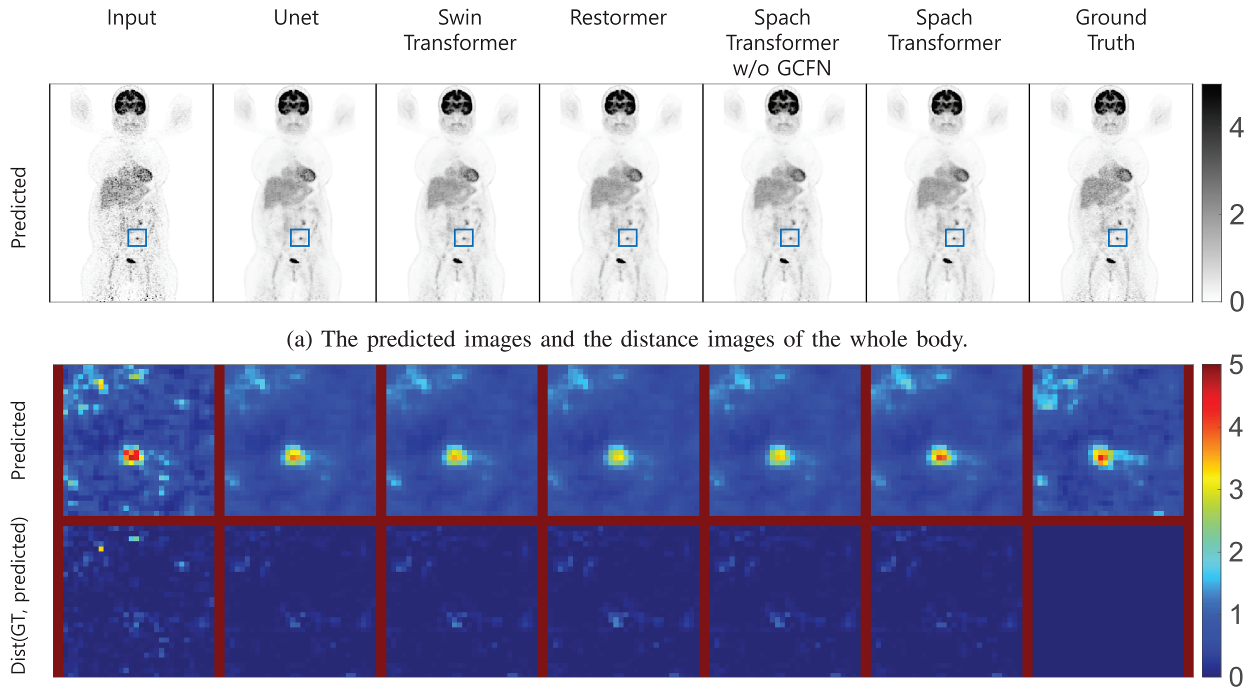
Fig. 6, 7, 8, and 9 present the coronal view of the results generated by the Spach Transformer and the reference methods



(a) The predicted images and the distance images of the whole body.

(b) The enlarged images where the blue rectangles are located in (a).

Fig. 6: Coronal and zoomed-in views of the denoised PET results of different methods based on an ^{18}F-DCFPyL dataset

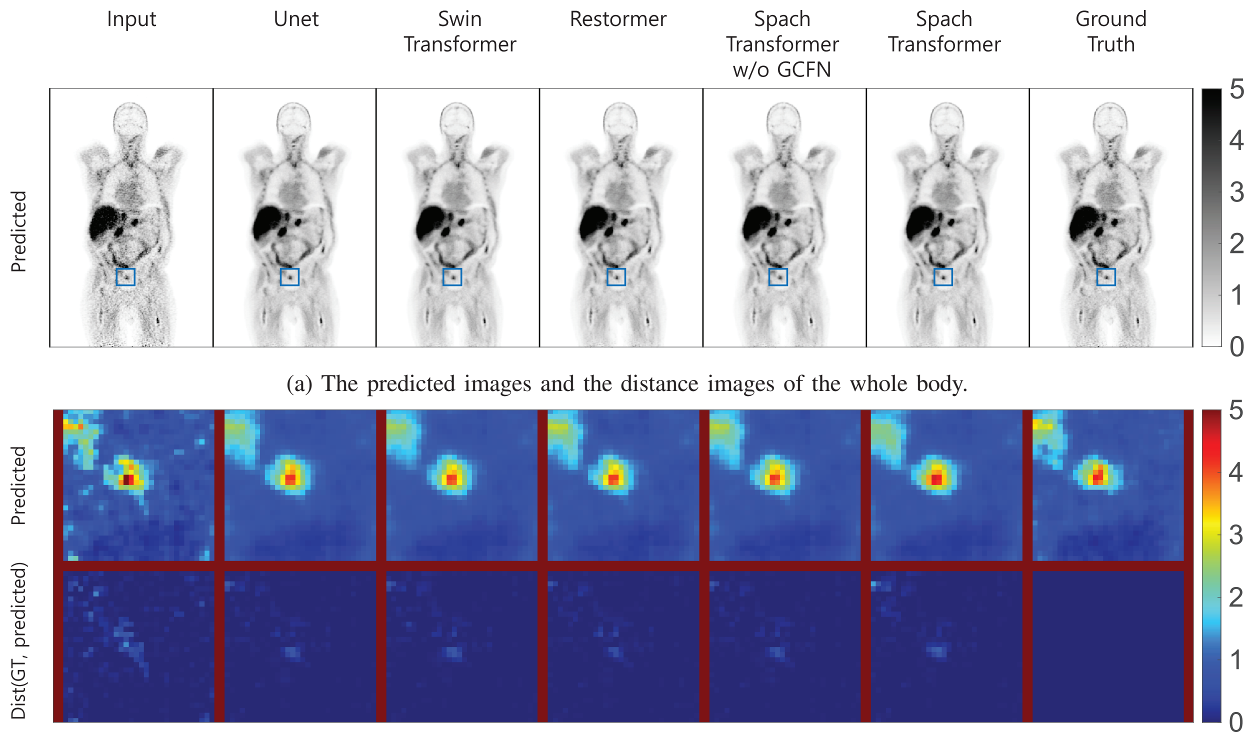


(a) The predicted images and the distance images of the whole body.

(b) The enlarged images where the blue rectangles are located in (a).

Fig. 7: Coronal and zoomed-in views of the denoised PET results of the state-of-the-art methods based on an ^{18}F-FDG dataset

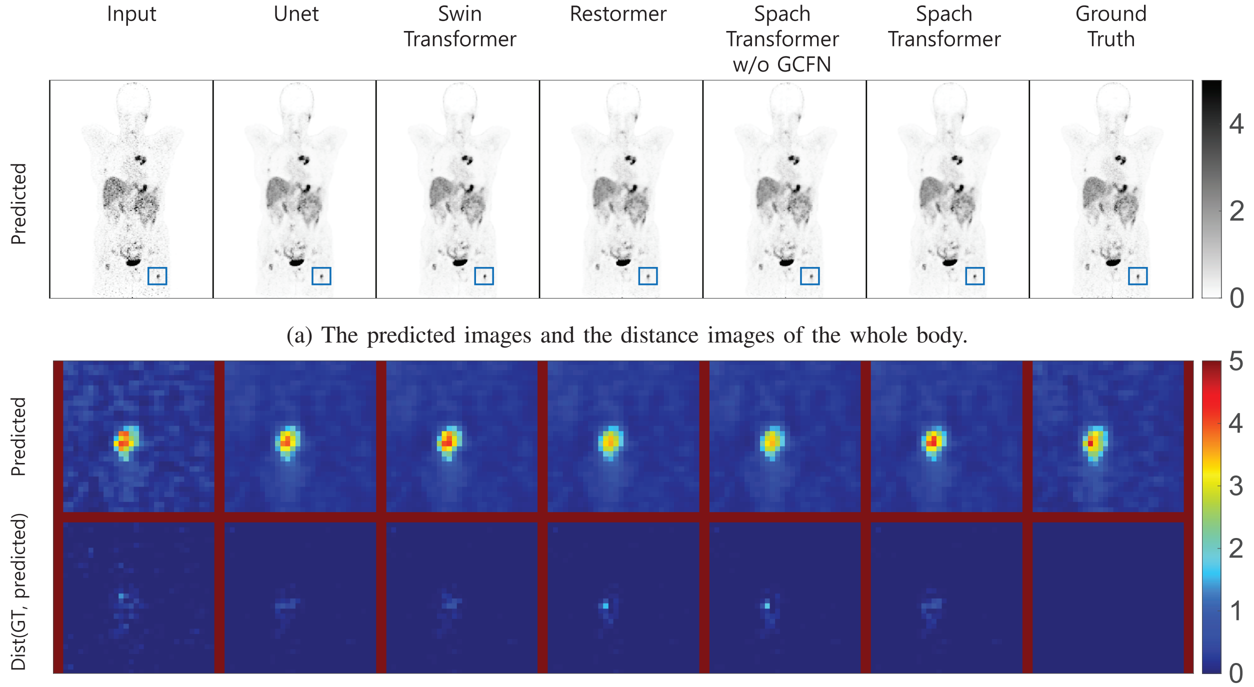
based on one ^{18}F-DCFPyL, one ^{18}F-FDG, one ^{18}F-ACBC, and one ^{68}Ga DOTATATE dataset. From Fig. 6a, 7a, 8a, and 9a, it can be observed that Restormer has better denoising performance than that of the Unet and Swin Transformer. It can also be observed that the Unet and Swin Transformer can achieve relatively higher uptake values than that of the Restormer in the tumor regions. This might be because the Unet and Swin Transformer focus more on spatial information, whereas the Restormer is highly dependent on the channel information. As the Spach Transformer can utilize both spatial and channel



(a) The predicted images and the distance images of the whole body.

(b) The enlarged images where the blue rectangles are located in (a).

Fig. 8: Coronal and zoomed-in views of the denoised PET results of different methods based on an ^{18}F-ACBC dataset



(a) The predicted images and the distance images of the whole body.

(b) The enlarged images where the blue rectangles are located in (a).

Fig. 9: Coronal and zoomed-in views of the denoised PET results of different methods based on a ^{68}Ga DOTATATE dataset

information together, it can better preserve the tumor details while also achieving competitive denoising performance than other reference methods.

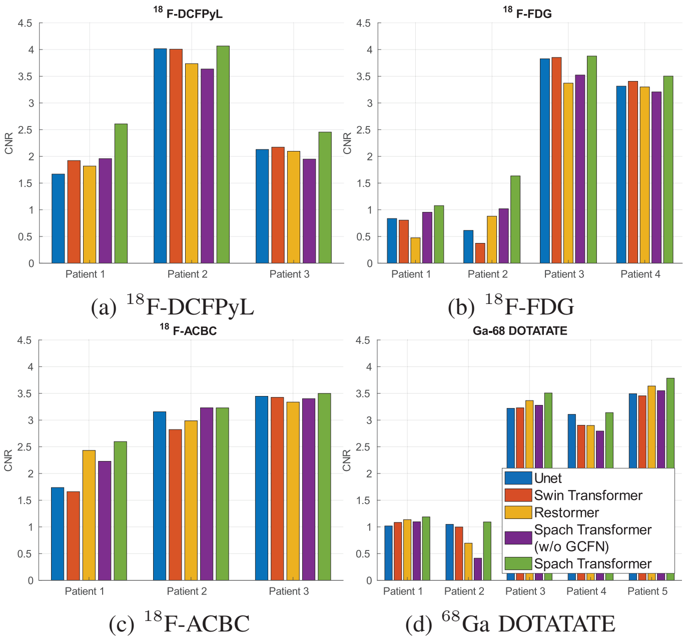
Table II and III show that the Spach Transformer achieved average PSNR of 55.0840 and aveerage SSIM of 0.9403, better than the Unet, Swin Transformer, Restormer, and Spach Transformer without GCFN. Compared to Restormer, Spach Transformer achieved a PSNR gain of 0.305 and an SSIM gain of 0.008 because of the extra spatial information utilized. Compared to Swin Transformer, Spach Transformer achieved

TABLE II: Quantitative comparison of PSNR with the state-of-the-art methods.

Tracer	Unet		Swin Transformer		Restormer		Spach Transformer w/o GCFN		Spach Transformer	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
^{18}F-DCFPyL	53.9726	5.7870	53.5862	5.7802	53.8871	5.7728	53.9818	5.7754	**54.0927**	**5.7879**
^{18}F-FDG	52.6028	3.1194	52.1079	3.2607	52.5813	3.0841	52.6135	3.0513	**52.8678**	**2.9585**
^{18}F-ACBC	58.3996	7.0715	57.6852	7.2052	58.4845	7.0278	58.6009	7.0026	**58.7500**	**6.9883**
^{68}Ga DOTATATE	54.0784	3.7722	53.3199	3.8628	54.4369	3.7322	54.4289	3.7438	**54.8294**	**3.6679**
Avg	54.6492	5.3646	54.0323	5.4194	54.7790	5.3398	54.8267	5.3439	**55.0840**	**5.2985**
Model Parameters	21,781,520		20,631,864		21,156,805		21,022,921		19,218,323	

TABLE III: The SSIM values of different methods on ^{18}F-DCFPyL, ^{18}F-FDG, ^{18}F-ACBC and ^{68}Ga DOTATATE datasets.

Tracer	Unet		Swin Transformer		Restormer		Spach Transformer w/o GCFN		Spach Transformer	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
^{18}F-DCFPyL	0.9234	0.0076	0.9125	0.0080	0.9370	0.0078	0.9369	0.0079	**0.9439**	**0.0069**
^{18}F-FDG	0.9364	0.0063	0.9282	0.0079	0.9472	0.0075	0.9477	0.0076	**0.9539**	**0.0074**
^{18}F-ACBC	0.8894	0.0059	0.8759	0.0052	0.8955	0.0061	0.8975	0.0059	**0.9138**	**0.0048**
^{68}Ga-DOTATATE	0.9121	0.0168	0.9166	0.0122	0.9432	0.0119	0.9417	0.0129	**0.9458**	**0.0127**
Avg	0.9148	0.0196	0.9097	0.0204	0.9328	0.0215	0.9327	0.0208	**0.9404**	**0.0169**



(a) ^{18}F-DCFPyL

(b) ^{18}F-FDG

(c) ^{18}F-ACBC

(d) ^{68}Ga DOTATATE

Fig. 10: The CNR values of different methods based on (a) ^{18}F-DCFPyL, (b) ^{18}F-FDG, (c) ^{18}F-ACBC and (d) ^{68}Ga DOTATATE datasets.

a PSNR gain of 1.052 and an SSIM gain of 0.031 because of the global channel information leveraged.

To examine the local performance of different methods at the tumor regions, we have further quantified the tumor CNR for datasets with confirmed tumors. Fig. 10 shows the CNR results of all the methods for datasets with different tracers. As observed from Fig. 6, 7, 8, and 9, the Restormer has worse performance regarding tumor-uptake preserving, which is confirmed by the CNR plot in Fig. 10. The Spach Transformer obtained better CNR results in the tumor regions than other reference methods because of the combination of spatial and channel information.

V. Discussion

For PET image denoising, one challenge is to reduce the noise while also preserving the tumor uptake. The latest deep learning architectures (e.g., the Swin Transformer and the Restormer) have focused on utilizing spatial or channel information in an efficient way to improve the specific image-processing task. In this work, we proposed a novel spatial and channel-wise transformer architecture, Spach Transformer, aiming to suppress the noise and at the same time preserving the image details. Oncology PET datasets from different tracers were utilized to evaluate the performance of different network structures.

From the evaluations, we observed that the spatial information-oriented network architectures (e.g., the Unet and Swin Transformer) can highlight the tumor region well but the results are still noisy. In contrast, the channel information-oriented network architecture (e.g., the Restormer) can achieve better denoising performance but failed to preserve the tumor uptakes, which can be observed through the CNR plot and the zoomed-in tumor regions of the coronal views. The SSIM, PSNR and CNR quantification show that the proposed Spach Transformer can achieve the best performance, demonstrating the benefits of combining both spatial and channel information.

Another thing to note is that all the network structures were trained using ^{18}F-FDG and ^{18}F-ACBC datasets, while the testing datasets include two additional tracers, i.e., ^{18}F-DCFPyL and ^{68}Ga DOTATATE, which were unseen during the training phase. ^{18}F-DCFPyL is a newly FDA approved prostate-specific membrane antigen (PSMA) tracer for prostate cancer diagnosis, which does not have many cases available when we began this work. ^{18}Ga-DOTATATE is for neuroendocrine tumor diagnosis and has worse image quality as compared to the ^{18}F-based tracer. Evaluations on datasets from these two tracers can help evaluate the robustness of the network to unseen data distributions and explore the possibility of deep learning-based cross-tracer applications. Results show that all the networks have good performance on those unseen data distributions and the proposed Spach Transformer has the best performance. Apart from cross-tracer testing, further evaluating the performance of the networks on datasets from different scanners is one of our future tasks.

VI. Conclusion

In this work, we proposed a spatial and channel-wise transformer that utilized the spatial and channel information for PET image denoising. The proposed Spach Transformer achieved better performances in terms of CNR, PSNR, and SSIM on ^{18}F-FDG, ^{18}F-ACBC, ^{18}F-DCFPyL, and ^{68}Ga-DOTATATE datasets compared to other reference methods. Our future work will focus on further quantitative evaluations based more clinical datasets.

References

[1] W. H. Sweet, "The uses of nuclear disintegration in the diagnosis and treatment of brain tumor," *New England Journal of Medicine*, vol. 245, no. 23, pp. 875–878, 1951.

[2] S. R. Cherry, T. Jones, J. S. Karp *et al.*, "Total-body pet: maximizing sensitivity to create new opportunities for clinical research and patient care," *Journal of Nuclear Medicine*, vol. 59, no. 1, pp. 3–12, 2018.

[3] J. W. Fletcher, B. Djulbegovic, H. P. Soares *et al.*, "Recommendations on the use of 18f-fdg pet in oncology," *Journal of Nuclear Medicine*, vol. 49, no. 3, pp. 480–508, 2008.

[4] T. Beyer, D. W. Townsend, T. Brun *et al.*, "A combined pet/ct scanner for clinical oncology," *Journal of nuclear medicine*, vol. 41, no. 8, pp. 1369–1379, 2000.

[5] M. F. Di Carli, S. Dorbala, J. Meserve *et al.*, "Clinical myocardial perfusion pet/ct," *Journal of Nuclear Medicine*, vol. 48, no. 5, pp. 783–793, 2007.

[6] G. El Fakhri, S. Surti, C. M. Trott *et al.*, "Improvement in lesion detection with whole-body oncologic time-of-flight pet," *Journal of Nuclear Medicine*, vol. 52, no. 3, pp. 347–353, 2011.

[7] J. Dutta, R. M. Leahy, and Q. Li, "Non-local means denoising of dynamic pet images," *PloS one*, vol. 8, no. 12, p. e81390, 2013.

[8] N. Boussion, C. Cheze Le Rest, M. Hatt *et al.*, "Incorporation of wavelet-based denoising in iterative deconvolution for partial volume correction in whole-body pet imaging," *European journal of nuclear medicine and molecular imaging*, vol. 36, no. 7, pp. 1064–1075, 2009.

[9] B. T. Christian, N. T. Vandehey, J. M. Floberg *et al.*, "Dynamic pet denoising with hypr processing," *Journal of Nuclear Medicine*, vol. 51, no. 7, pp. 1147–1154, 2010.

[10] C. W. Bevington, J.-C. Cheng, and V. Sossi, "A 4-D iterative HYPR denoising operator improves PET image quality," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 6, no. 6, pp. 641–655, 2021.

[11] J. Yan, J. C.-S. Lim, and D. W. Townsend, "Mri-guided brain pet image filtering and partial volume correction," *Physics in medicine & biology*, vol. 60, no. 3, p. 961, 2015.

[12] L. Xiang, Y. Qiao, D. Nie *et al.*, "Deep auto-context convolutional neural networks for standard-dose pet image estimation from low-dose pet/mri," *Neurocomputing*, vol. 267, pp. 406–416, 2017.

[13] K. Gong, J. Guan, C.-C. Liu *et al.*, "Pet image denoising using a deep neural network through fine tuning," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 3, no. 2, pp. 153–161, 2018.

[14] S. Kaplan and Y.-M. Zhu, "Full-dose pet image estimation from low-dose pet image using deep learning: a pilot study," *Journal of digital imaging*, vol. 32, no. 5, pp. 773–778, 2019.

[15] J. Cui, K. Gong, N. Guo *et al.*, "Pet image denoising using unsupervised deep learning," *European journal of nuclear medicine and molecular imaging*, vol. 46, no. 13, pp. 2780–2789, 2019.

[16] K. T. Chen, E. Gong, F. B. de Carvalho Macruz *et al.*, "Ultra–low-dose 18f-florbetaben amyloid pet imaging using deep learning with multi-contrast mri inputs," *Radiology*, vol. 290, no. 3, pp. 649–656, 2019.

[17] F. Hashimoto, H. Ohba, K. Ote *et al.*, "Dynamic pet image denoising using deep convolutional neural networks without prior training datasets," *IEEE access*, vol. 7, pp. 96 594–96 603, 2019.

[18] C. O. da Costa-Luis and A. J. Reader, "Micro-networks for robust mr-guided low count pet imaging," *IEEE transactions on radiation and plasma medical sciences*, vol. 5, no. 2, pp. 202–212, 2020.

[19] G. Schramm, D. Rigie, T. Vahle *et al.*, "Approximating anatomically-guided pet reconstruction in image space using a convolutional neural network," *Neuroimage*, vol. 224, p. 117399, 2021.

[20] A. Mehranian, S. D. Wollenweber, M. D. Walker *et al.*, "Image enhancement of whole-body oncology [18f]-fdg pet scans using deep neural networks to reduce noise," *European journal of nuclear medicine and molecular imaging*, vol. 49, no. 2, pp. 539–549, 2022.

[21] R. S. Daveau, I. Law, O. M. Henriksen *et al.*, "Deep learning based low-activity pet reconstruction of [11c] pib and [18f] fe-pe2i in neurodegenerative disorders," *NeuroImage*, vol. 259, p. 119412, 2022.

[22] K. Kim, D. Wu, K. Gong *et al.*, "Penalized pet reconstruction using deep learning prior and local linear fitting," *IEEE transactions on medical imaging*, vol. 37, no. 6, pp. 1478–1487, 2018.

[23] K. Gong, J. Guan, K. Kim *et al.*, "Iterative pet image reconstruction using convolutional neural network representation," *IEEE transactions on medical imaging*, vol. 38, no. 3, pp. 675–685, 2018.

[24] K. Gong, C. Catana, J. Qi *et al.*, "Pet image reconstruction using deep image prior," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1655–1665, 2018.

[25] I. Häggström, C. R. Schmidtlein, G. Campanella *et al.*, "Deeppet: A deep encoder–decoder network for directly solving the pet image reconstruction inverse problem," *Medical image analysis*, vol. 54, pp. 253–262, 2019.

[26] H. Lim, I. Y. Chun, Y. K. Dewaraja *et al.*, "Improved low-count quantitative pet reconstruction with an iterative neural network," *IEEE transactions on medical imaging*, vol. 39, no. 11, pp. 3512–3522, 2020.

[27] A. Mehranian and A. J. Reader, "Model-based deep learning pet image reconstruction using forward–backward splitting expectation–maximization," *IEEE transactions on radiation and plasma medical sciences*, vol. 5, no. 1, pp. 54–64, 2020.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[29] J. Ouyang, K. T. Chen, E. Gong *et al.*, "Ultra-low-dose pet reconstruction using generative adversarial network with feature matching and task-specific perceptual loss," *Medical physics*, vol. 46, no. 8, pp. 3555–3564, 2019.

[30] Z. Hu, H. Xue, Q. Zhang *et al.*, "Dpir-net: Direct pet image reconstruction based on the wasserstein generative adversarial network," *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 5, no. 1, pp. 35–43, 2020.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[32] Y. Lei, X. Dong, T. Wang *et al.*, "Whole-body pet estimation from low count statistics using cycle-consistent generative adversarial networks," *Physics in Medicine & Biology*, vol. 64, no. 21, p. 215017, 2019.

[33] L. Zhou, J. D. Schaefferkoetter, I. W. Tham *et al.*, "Supervised learning with cyclegan for low-dose fdg pet image denoising," *Medical image analysis*, vol. 65, p. 101770, 2020.

[34] T.-A. Song, S. R. Chowdhury, F. Yang *et al.*, "Pet image super-resolution using generative adversarial networks," *Neural Networks*, vol. 125, pp. 83–91, 2020.

[35] A. Sanaat, I. Shiri, H. Arabi *et al.*, "Deep learning-assisted ultra-fast/low-dose whole-body pet/ct imaging," *European journal of nuclear medicine and molecular imaging*, vol. 48, no. 8, pp. 2405–2415, 2021.

[36] S. Xue, R. Guo, K. P. Bohn *et al.*, "A cross-scanner and cross-tracer deep learning method for the recovery of standard-dose imaging quality from low-dose pet," *European journal of nuclear medicine and molecular imaging*, vol. 49, no. 6, pp. 1843–1856, 2022.

[37] A. Vaswani, N. Shazeer, N. Parmar *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] K. Han, Y. Wang, H. Chen *et al.*, “A survey on vision transformer,” *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[40] Z. Liu, Y. Lin, Y. Cao *et al.*, “Swin transformer: Hierarchical vision transformer using shifted windows,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[41] S. W. Zamir, A. Arora, S. Khan *et al.*, “Restormer: Efficient transformer for high-resolution image restoration,” *arXiv preprint arXiv:2111.09881*, 2021.

[42] J. Chen, Y. Lu, Q. Yu *et al.*, “Transunet: Transformers make strong encoders for medical image segmentation,” *arXiv preprint arXiv:2102.04306*, 2021.

[43] Y. Xie, J. Zhang, C. Shen *et al.*, “Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 171–180.

[44] W. Wang, C. Chen, M. Ding *et al.*, “Transbts: Multimodal brain tumor segmentation using transformer,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 109–119.

[45] Y. Zhang, H. Liu, and Q. Hu, “Transfuse: Fusing transformers and cnns for medical image segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 14–24.

[46] A. Hatamizadeh, Y. Tang, V. Nath *et al.*, “Unetr: Transformers for 3d medical image segmentation,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[47] J. Chen, Y. Du, Y. He *et al.*, “Transmorph: Transformer for unsupervised medical image registration,” *arXiv preprint arXiv:2111.10480*, 2021.

[48] D. Hendrycks and K. Gimpel, “Gaussian error linear units (gelus),” *arXiv preprint arXiv:1606.08415*, 2016.

[49] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer normalization,” *arXiv preprint arXiv:1607.06450*, 2016.

[50] Y. N. Dauphin, A. Fan, M. Auli *et al.*, “Language modeling with gated convolutional networks,” in *International conference on machine learning*. PMLR, 2017, pp. 933–941.

[51] J. T. Barron, “A general and adaptive robust loss function,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4331–4339.

[52] I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” *arXiv preprint arXiv:1608.03983*, 2016.