# MAK-GAN: Multi-level Adaptive Convolutional Kernels for Asymmetric Multi-modal PET Reconstruction

Xinyi Zeng[1], Pinxian Zeng[1], Yan Wang[1] (✉), Jiaqi Cui[1], Luping Zhou[2], Caiwen Jiang[3], Han Zhang[3], Dinggang Shen[3,4,5]

[1] School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] School of Electrical and Information Engineering, University of Sydney, Australia
[3] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, Shanghai Tech University, Shanghai, China
[4] Shanghai United Imaging Intelligence Co., Ltd., Shanghai, China
[5] Shanghai Clinical Research and Trial Center, Shanghai, China

**Abstract.** Positron emission tomography (PET) reconstruction from low-dose to standard-dose acquisitions poses a significant challenge in medical imaging. While integrating Magnetic Resonance Imaging (MRI) for complementary guidance shows promise for enhancing reconstruction fidelity, current multi-modal approaches typically treat PET and MRI uniformly, neglecting their inherent asymmetry within the multi-modal context. This leads to insufficient utilization of anatomical guidance provided by MRI and neglects the unique metabolic characteristics of PET. To address these limitations, we propose MAK-GAN, a novel **G**enerative **A**dversarial **N**etwork (GAN) that incorporates **M**ulti-level **A**daptive **K**ernels to distinguish feature extraction and interaction strategies between the primary (PET) and auxiliary (MRI) modalities in the asymmetric multi-modal PET reconstruction task. Specifically, we design a Multi-Kernel Extraction (MKE) block for both PET and MRI branches, replacing linear projections in vanilla Transformers with hierarchical multi-kernel convolutions. This enables efficient extraction of modality-specific features at multiple scales while reducing computational overhead. Subsequently, we asymmetrically introduce an Adaptive-Kernel Interaction (AKI) block in the PET branch. This block integrates self- and cross-attention modules to dynamically generate weights for adaptive kernels, preserving PET-specific characteristics while utilizing MRI's anatomical information. Finally, we incorporate two PET-centric optimization strategies to prioritize PET during reconstruction: a residual connection for direct LPET-to-SPET mapping, and an edge-aware consistency loss to enforce structural coherence. Experiments demonstrate superiority on two PET/MRI datasets.

**Keywords:** Asymmetric Multi-modal, PET Reconstruction, Adaptive Convolutional Kernels, Feature Extraction and Fusion.

The first two authors (Xinyi Zeng and Pinxian Zeng) contribute equally to this work.

# 1 Introduction

Positron emission tomography (PET) is a pivotal diagnostic tool for non-invasive physiological imaging [1]. While the standard-dose PET (SPET) provides high diagnostic clarity, its reliance on high radioactive tracer doses raises safety concerns [2-4]. Low-dose PET (LPET) presents a safer alternative but compromises image quality. Consequently, reconstructing high-fidelity SPET from LPET acquisitions has become critical, aiming to balance radiation safety with the retention of clinically relevant details.

Thanks to deep learning (DL), the development of medical image analysis has made great progress [5-16]. In PET reconstruction, convolutional neural networks (CNNs) initially led the field, with U-Net-based architectures commonly used to build the mapping from LPET to SPET. For example, Wang *et al.* [6] first introduced the Generative Adversarial Network (GAN) with conditional mechanisms to enhance high-frequency details. However, CNN-based methods struggle to model long-range metabolic dependencies due to their focus on local receptive fields. To overcome this, hybrid architectures combining Vision Transformers (ViTs) [11] have emerged, adopting self-attention mechanisms to capture global correlations. Luo *et al.* [12] added stacked Transformer layers as intermediate bottleneck layers between the CNN encoder and decoder. Despite these advancements, Transformer-based methods face challenges with quadratic complexity, limiting their application to small-scale features or requiring compromises like single-scale convolutional projections [17]. These limitations hinder multi-scale feature extraction, affecting the recovery of fine-grained anatomical structures.

Recent efforts have expanded into multi-modal paradigms that integrate complementary information from auxiliary imaging modalities, particularly Magnetic Resonance Imaging (MRI) [18-24]. Conventional methods typically concatenate MRI and PET along the channel dimension, treating them as a unified input for end-to-end reconstruction [19]. Nevertheless, such strategies fail to explicitly model cross-modal interactions, overlooking the distinct representations encoded in each modality. Some multi-modal methods address this by using identical encoders for feature extraction, followed by simple fusion techniques (e.g., summation or multiplication) to merge the modality-specific features [20]. While these methods work well in "*equivalent multi-modal*" scenarios, such as multi-parametric MRI segmentation, they are less effective in "*asymmetric multi-modal*" settings like multi-modal PET-MRI reconstruction. In such cases, LPET should be prioritized as the primary modality due to its direct correlation with the target SPET image in both metabolic mapping and structural continuity, whereas MRI should act as a complementary modality to provide additional anatomical guidance. Existing multi-modal PET reconstruction approaches, however, uniformly process both modalities with identical encoder architectures and fusion rules, which, in such an asymmetric multi-modal setting, undermines the critical role of LPET while underutilizing the contextual cues provided by MRI.

Recent advancements in dynamic convolutional kernels have shown strong potential in natural image classification by enabling context-aware feature extraction through adaptive weight modulation [25-29]. However, such techniques remain underexplored in medical imaging, let alone volumetric dense prediction tasks that require modeling asymmetric cross-modal interactions. In this paper, we propose MAK-GAN, a novel

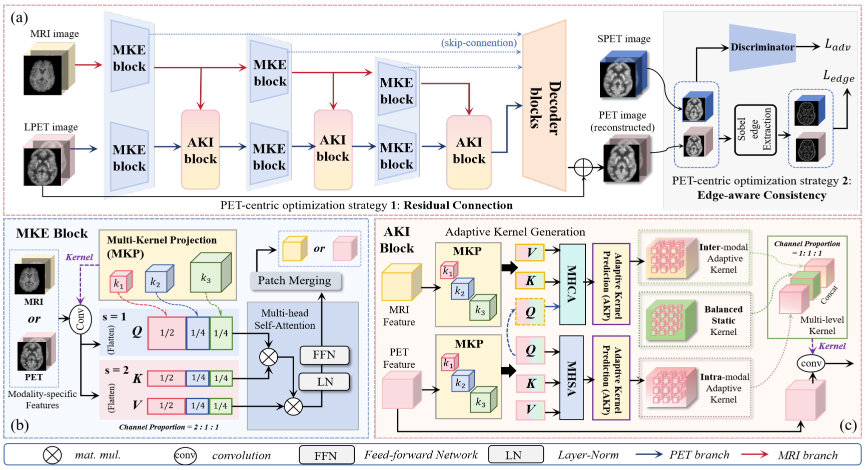

**Fig. 1.** Overview of our MAK-GAN model: (a) Two PET-centric optimization strategies, (b) MKE utilizes multi-kernel convolutional projections, and (c) AKI with self- and cross-attention mechanisms for adaptive kernel prediction.

**GAN**-based framework that integrates **M**ulti-level **A**daptive convolutional **K**ernels to explicitly differentiate feature extraction and interaction strategies between primary (LPET) and auxiliary (MRI) modalities in asymmetric multi-modal settings. Specifically, we hierarchically build Multi-Kernel Extraction (MKE) blocks for both PET and MRI branches. By constructing multi-kernel projection with varying kernel sizes, the MKE block enhances multi-scale modality-specific feature extraction while reducing computational overhead. Furthermore, to capture the primary-auxiliary relationship in PET-MRI interaction, we introduce the Adaptive-Kernel Interaction (AKI) block in the PET branch. The AKI block employs parallel self- and cross-attention modules to dynamically generate weights for intra- and inter-modal adaptive kernels, which are combined with a balanced static kernel. This multi-level kernel ensures that the interacted feature preserves sufficient PET-specific metabolic characteristics while adaptively leveraging MRI-derived anatomical information. Finally, we employ two PET-centric optimization strategies to further reinforce the dominance of PET: a residual connection and an edge-aware consistency loss. Our contributions are as follows:

1) We propose MAK-GAN, a novel framework that integrates multi-level adaptive kernels to differentiate feature extraction and interaction strategies between primary (LPET) and auxiliary (MRI) modalities, addressing the inherent asymmetry in multi-modal PET reconstruction tasks.
2) Multi-Kernel Extraction (MKE) block is introduced for both PET and MRI branches, replacing linear projections with hierarchical multi-kernel convolutions, enabling efficient extraction of multi-scale modality-specific features.
3) Adaptive-Kernel Interaction (AKI) block is asymmetrically incorporated into the PET branch, which integrates self- and cross-attention to preserve PET-specific characteristics while adaptively leveraging MRI's anatomical information.
4) We employ two PET-centric optimization strategies to prioritize PET during reconstruction. Experimental results demonstrate the superiority of our method.

## 2 Methodology

As shown in Fig. 1, our MAK-GAN is based on a Generative Adversarial Network. The generator takes aligned LPET images ($I_L$) and corresponding MRI ($I_M$) to estimate high-quality PET images ($I_E$), aiming to match the target SPET ($I_S$). The discriminator distinguishes between authentic $\{I_S,\ I_L\}$ pairs and synthesized pairs $\{I_E,\ I_L\}$. The objective of the generator is to fool the discriminator, forming a min-max game.

Notably, our MAK-GAN focuses on the encoder of the generator, incorporating two core blocks at each stage. The Multi-Kernel Extraction (MKE) block uses multi-kernel projections with varying kernel sizes to capture modality-specific features for both PET and MRI. These features are then fed into the Adaptive-Kernel Interaction (AKI) Block, asymmetrically embedded in the PET branch, where two adaptively generated convolution kernels are balanced with a static kernel to form robust cross-modal interaction.

### 2.1 Multi-Kernel Extraction Block

Unlike conventional methods limited to small-scale transformer-based processing [12] or single-scale convolutional projections embedded in Transformers [13], we propose a Multi-Kernel Extraction (MKE) block for both MRI and PET branches. This block incorporates the Multi-Kernel Projection (MKP) strategy into self-attention mechanism. By replacing linear projections with hierarchical multi-kernel convolutions, our MKP enables efficient multi-scale feature extraction with global contexts, enhancing modality-specific representation while optimizing computational efficiency.

Specifically, in the $i$-th stage ($i \in [1,3]$), the MKE block takes modality-specific feature $F_i \in R^{C_i \times H_i \times W_i \times D_i}$ as input, where $H_i$, $W_i$, and $D_i$ represent height, width, and depth, and $C_i$ denotes the number of channels. For $i = 1$, $F_i$ corresponds to the image $I_L$ or $I_M$. As shown in Fig.1 (b), the MKP strategies ($MKP(\cdot)$) applies three convolutions $\{k_1(\cdot), k_2(\cdot), k_3(\cdot)\}$ with varying kernel sizes of 3, 5, and 7 to derive multi-scale components. These components are concatenated along the channel dimension to form the query $Q$, key $K$, and value $V$. The process is expressed as:

$$Q, K, V = MKP(F_i), \text{ where } Q = Flat(Cat(k_1(F_i), k_2(F_i), k_3(F_i))), \tag{1}$$

$$K, V = Flat(Cat(k_1(F_i, s = 2), k_2(F_i, s = 2), k_3(F_i, s = 2))), \tag{2}$$

where $Flat(\cdot)$ represents the flattening operation and $Cat(\cdot)$ denotes concatenation. Notably, different strides $s$ are employed for convolution operations to reduce computational costs with minimal performance loss ($s = 1$ for $Q$, while $s = 2$ for $K$ and $V$). Additionally, paddings are adjusted to ensure consistent spatial sizes of the projection components. Once the projections are obtained, multi-head self-attention (MHSA) [11], denoted as $MHSA(\cdot)$, is applied to explore the long-range dependencies:

$$MHSA(Q, K, V) = Cat(head_1, \ldots, head_h), \tag{3}$$

$$head_j = Attn(Q, K, V) = Softmax(Q \cdot K/\sqrt{c}) \cdot V, \tag{4}$$

where each head $head_j$ computes the attention mechanism $Attn(\cdot)$. Here, $j \in [1, h]$ with the number of heads $h$ set to 4. Subsequently, Layer-normalization (LN), a Feed-forward Network (FFN), and a patch-merging operation are applied to obtain the output of the MKE, denoted as $O_i \in R^{C_i' \times H_i' \times W_i' \times D_i'}$, where $H_i' \times W_i' \times D_i' = \frac{H_i}{2} \times \frac{W_i}{2} \times \frac{D_i}{2}$, and $C_1' = 64$, which doubles at each stage. Notably, $F_i$ and $O_i$ are denoted as $F_i^{PET}$ and $F_i^{MRI}$, and $O_i^{PET}$ and $O_i^{MRI}$ for the PET and MRI branches, respectively. The above workflow in the MKE block preserves unique modality-specific characteristics before cross-modal interaction.

## 2.2 Adaptive-Kernel Interaction Block

Leveraging dynamic convolution techniques [25-27], which enable context-aware adaptation through weight modulation, we propose the Adaptive-Kernel Interaction (AKI) block to model primary-auxiliary relationships in asymmetric multi-modal PET reconstruction. As shown in Fig. 1 (c), the AKI block is integrated into the PET branch, dynamically calibrating the influence of MRI anatomical information on PET feature refinement. It comprises three components: the Intra-modal Adaptive Kernel, Inter-modal Adaptive Kernel, and Balanced Static Kernel, which together form a multi-level filtering mechanism to refine PET features across varying levels of abstraction.

**Intra-modal Adaptive Kernel (Intra-AK).** The Intra-AK aims to enhance the PET features extracted by the MKE block, ensuring that the rich functional information in the PET data is fully preserved and effectively utilized during the interaction process. Specifically, given the PET feature $O_i^{PET}$ from the MKE block at the $i$-th stage, we first apply the multi-kernel projection and self-attention calculation, as in the MKE block, to derive the intermediate kernel-prior feature $f_i^{Intra}$. An adaptive kernel predictor (AKP) is then introduced, utilizing the $f_i^{Intra}$ to dynamically generate the convolutional kernel weight $\theta_i^{Intra}$ for Intra-AK. The process can be formulated as:

$$f_i^{Intra} \in R^{C_i' \times H_i' \times W_i' \times D_i'} = MHSA(Q, K, V = MKP(O_i^{PET})), \tag{5}$$

$$\theta_i^{Intra} \in R^{C_i^{Intra} \times C_i' \times (k_a \times k_a \times k_a)} = AKP(f_i^{Intra}), \tag{6}$$

where $MKP(\cdot)$ denotes the multi-kernel projection in the MKE block, and $AKP(\cdot)$ represents the adaptive kernel prediction, comprising multiple linear layers and a reshape operation. $C_i^{Intra}$ and $C_i'$ represent the output and input channels, respectively, and $k_a$ is a predefined kernel size for the adaptive kernels.

**Inter-modal Adaptive Kernel (Inter-AK).** To fully leverage the auxiliary MRI modality and dynamically assess its influence on PET reconstruction, we propose the Inter-AK. This adaptive kernel facilitates effective inter-modal interaction by integrating outputs from the MKE blocks of both PET and MRI branches, $O_i^{PET}$ and $O_i^{MRI}$.

Multi-kernel projection $MKP(\cdot)$ and cross-attention (denoted as $MHCA(\cdot)$) are applied to generate the kernel-prior feature $f_i^{Inter}$, which is processed by the adaptive kernel predictor $AKP(\cdot)$ to generate dynamic convolutional kernel weight $\theta_i^{Inter}$ with the output channel dimension $C_i^{Inter}$. The operations can be expressed as:

$$f_i^{Inter} \in R^{C_i' \times H_i' \times W_i' \times D_i'} = MHCA(Q = MKP(O_i^{PET}), K, V = MKP(O_i^{MRI})), \tag{7}$$

$$\theta_i^{Inter} \in R^{C_i^{Inter} \times C_i' \times (k_a \times k_a \times k_a)} = AKP(f_i^{Inter}). \tag{8}$$

Note that, in the cross-attention calculation, $K$ and $V$ are derived from the MRI branch, while $Q$ is derived from the PET branch to query the most relevant MRI-derived anatomical information for PET reconstruction.

**Balanced Static Kernel (BSK).** To complement the two adaptive kernels, we also introduce the BSK to mitigate over-reliance on dynamic convolutions by preserving crucial basic information that might be significantly affected by the variability of adaptive weights, aiming to balance AKs by preventing their underlearning in early stages.

Unlike intra- and inter-AKs that adaptively determine weights based on features, the "static" nature in BSK lies in the use of randomly initialized learnable weight $\theta_i^B \in R^{C_i^B \times C_i' \times (k_a \times k_a \times k_a)}$ in conventional convolution with the same kernel size $k_a$, where $C_i^B$ denotes its output channel. The two AKs $\theta_i^{Intra}$ and $\theta_i^{Inter}$, along with the BSK $\theta_i^B$, are concatenated along the channel dimension to form a multi-level kernel $\theta_i^M$:

$$\theta_i^M \in R^{C_i' \times C_i' \times (k_a \times k_a \times k_a)} = Cat(\theta_i^{Intra}, \theta_i^{Inter}, \theta_i^B), \tag{9}$$

where $C_i' = C_i^{Intra} + C_i^{Inter} + C_i^{BS}$. Notably, the optimal ratio $\{C_i^{Intra}: C_i^{Inter}: C_i^B\}$ is explored and set as $\{1: 1: 1\}$. This equal channel allocation enhances the coordination of the three components: PET-specific self-calibration (via Intra-AK), cross-modal interaction (via Inter-AK), and stable feature preservation (via BSK). Finally, $\theta_i^M$ assigns three 3D convolutional kernels to filter each voxel in the initial PET feature $O_i^{PET}$:

$$F_{i+1} \in R^{C_i' \times H_i' \times W_i' \times D_i'} = O_i^{PET} * \theta_i^M, \tag{10}$$

where "$*$" denotes the convolution operator, and the filtered $F_{i+1}$ is for the next stage.

## 2.3 PET-centric Optimization Strategies

As illustrated in Fig. 1(a), we introduce two PET-centric strategies into the GAN-based architecture during the adversarial training process, aiming to reinforce the dominance of the PET modality and achieve more robust reconstruction results.

**End-to-end Residual Connection.** For the PET branch, in addition to the hierarchical encoder-decoder skip connections, we also establish an end-to-end residual connection (element-wise addition as "$\oplus$") directly from $I_L$ to the output $I_E$. This design constructs a more robust LPET-to-SPET mapping by enabling the model to focus on learning residual features rather than reconstructing the output from scratch, thereby reducing the difficulty of prediction and accelerating convergence during training [13].

**Edge-aware Consistency Loss.** In adversarial training, the generator $G$ synthesizes $I_E$, while the discriminator $D$ evaluates both real and synthesized image pairs, i.e., $\{I_S$, $I_L\}$ and $\{I_E, I_L\}$, to distinguish real from fake. Following prior work [6], the adversarial loss forms a min-max game, which can be defined as:

$$L_{adv} = E_{I_L, I_S}[log\, D(I_L, I_S)] + E_{I_L}[log(1 - D(I_L, I_E = G(I_L, I_M)))]. \tag{11}$$

Additionally, we introduce an edge-aware consistency loss to preserve structural details in the reconstructed PET images, guiding the model to prioritize the recovery of PET-related contextual information. Specifically, we apply a Sobel layer [30] to the image pair $\{I_S, I_E\}$ and then compute the L1 loss between the extracted edges:

$$L_{edge} = E[||Sobel(I_S) - Sobel(I_E)||]. \tag{12}$$

The final optimization objective can be represented as a combination of the adversarial loss, the edge-aware consistency loss, and the common L1 loss on $I_S$ and $I_E$:

$$L_{total} = L_{adv} + \alpha \cdot l_{edge} + \beta \cdot E[||I_S - I_E||], \tag{13}$$

where $\alpha$ and $\beta$ are two hyper-parameters used to balance the three terms.

## 3 Experiments and Results

**Datasets.** Our experiments utilize two in-house datasets. The Clinical dataset consists of PET brain images from 16 subjects (8 normal control (NC) and 8 with mild cognitive impairment (MCI)), acquired using a Siemens Biograph mMR PET-MR system. SPET scans were obtained within a 12-minute acquisition of tracer injection, while LPET simulations were captured in 3 minutes (25% standard dose). The Dynamic-PET dataset includes 18F-FDG PET/MR head images from another 16 subjects, acquired with a uPMR 790 PET/MR scanner with extreme low-dose protocols (4% standard dose).

**Implementation Details.** Our model is implemented in PyTorch and trained on an RTX 3090. Based on trial studies, the L1 loss (with $\alpha$) is slightly larger than the edge loss (with $\beta$) but much smaller than the adversarial loss. Following [12,20], optimal allocation of $\alpha = 100$ and $\beta = 50$ effectively balances these loss terms. The learning rates for $G$ and $D$ are set to $2 \times 10^{-4}$. The adaptive kernel size $k_a$ is set to 3. We employ the convolution blocks in U-Net [9] for the decoder. Both datasets undergo the same preprocessing: PET and MRI images are registered and resized to $128 \times 128 \times 128$, and each image is sliced into overlapping patches with a stride of 8, resulting in 729 patches with a size of 64 for each patient. Using patches as basic input units, the final PET images are formed by stitching estimated patches together and averaging overlapping regions. During training, leave-one-out cross-validation (LOOCV) is applied for 16 times, each with one subject for validation and the others for training.

**Table 1.** Comparison with five SPET reconstruction methods in terms of PSNR, SSIM, NMSE, and GFLOPS. * marks PSNR improvements with statistical significance.

| Type | Method | Clinical | | | Dynamic-PET | | | GFLOPs |
|---|---|---|---|---|---|---|---|---|
| | | PSNR | SSIM | NMSE | PSNR | SSIM | NMSE | |
| - | LPET | *21.068 | 0.977 | 0.0550 | *19.217 | 0.784 | 0.208 | - |
| single-modal | Ea-GAN [14] | *24.867 | 0.984 | 0.0235 | *22.735 | 0.830 | 0.130 | 70.38 |
| | 3D CVT-GAN [13] | *25.084 | 0.987- | 0.0218 | *22.604 | 0.836 | 0.135 | **23.80** |
| Multi-modal | M-Unet [9] | *24.442 | 0.984 | *0.0261 | *22.549 | 0.824 | 0.147 | 40.50 |
| | LA-GAN [20] | *24.807 | 0.984 | *0.0239 | *22.651 | 0.827 | 0.133 | 98.57 |
| | MTrans-GAN [21] | *25.106 | 0.987 | 0.0220 | 23.083 | 0.840 | 0.124 | 30.14 |
| | **MAK-GAN(ours)** | **25.311** | **0.988** | **0.0207** | **23.226** | **0.845** | **0.110** | 30.08 |

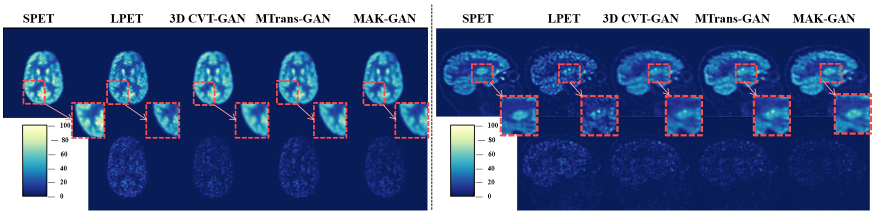

**Fig. 2.** Visualization results on Clinical dataset (left) and Dynamic-PET Dataset (right). The second row displays the error maps. The zoom-in areas are highlighted by red boxes.

**Comparison Analysis.** To demonstrate the superiority, we use SPET-corresponding LPET as the baseline and evaluate our proposed MAK-GAN against two single-modal methods (Ea-GAN [14], 3D CVT-GAN [13]) and three state-of-the-art multi-modal methods (M-Unet [9], LA-GAN [20], MTrans-GAN [21]). As shown in Table 1, our method outperforms all comparison methods across PSNR, SSIM, and NMSE. Specifically, on the Clinical dataset, MAK-GAN achieves a 0.205 PSNR improvement over the second-best performer MTrans-GAN. On the more challenging Dynamic-PET with ultra-low dose, our method enhances PSNR by 0.143 dB and reduces NMSE by 0.014, while maintaining reasonable computational consumption. This is attributed to our multi-kernel projection, which enables symmetric multi-scale feature extraction for both modalities and facilitates effective asymmetric cross-modal interaction. A paired t-test shows p-values for PSNR consistently below 0.05 for both datasets in most cases, confirming statistically significant improvements. Fig. 2 shows the visualization results, where our MAK-GAN provides the closest reconstruction result to the real SPET images with the smallest error compared to other methods.

**Table 2.** Quantitative comparison of ablation models on PSNR, SSIM and NMSE.

| Model | Description | Clinical | | | Dynamic-PET | | |
|---|---|---|---|---|---|---|---|
| | | PSNR | SSIM | NMSE | PSNR | SSIM | NMSE |
| A | Baseline | 24.084 | 0.980 | 0.0268 | 22.415 | 0.822 | 0.157 |
| B | A + MKE Module | 24.617 | 0.983 | 0.0240 | 22.569 | 0.830 | 0.140 |
| C | B + AKI Module | 24.934 | 0.986 | 0.0226 | 22.981 | 0.834 | 0.132 |
| D | C + Residual Connection | 25.169 | 0.987 | 0.0218 | 23.153 | 0.838 | 0.122 |
| **E (Proposed)** | D + Edge-aware Loss | **25.311** | **0.988** | **0.0207** | **23.226** | **0.845** | **0.110** |

**Ablation studies.** We conducted the following ablation experiments: (1) A GAN-based baseline using the standard U-Net encoder for PET and MRI branches, with simple convolution for fusion, and no edge-aware consistency loss or residual connection (Model-A). (2) Replacing the encoder blocks with MKE blocks for both MRI and PET branches (Model-B); (3) Adding AKI blocks to Model-B (Model-C); (4) Incorporating the residual connection (Model-D); (5) Introducing edge-aware loss to form our proposed model (Model-E). As shown in Table 2, the comparison of Model-A and Model-B reveals that the MKE block improves the performance through its multi-scale modality-specific feature extraction. Moving from Model-B to Model-C, the inclusion of the AKI block leads to better results, demonstrating its effectiveness in inter-modal interaction. Finally, the enhancements in the proposed Model-E demonstrate that the PET-centric optimization strategies effectively reinforce the dominance of the PET modality.

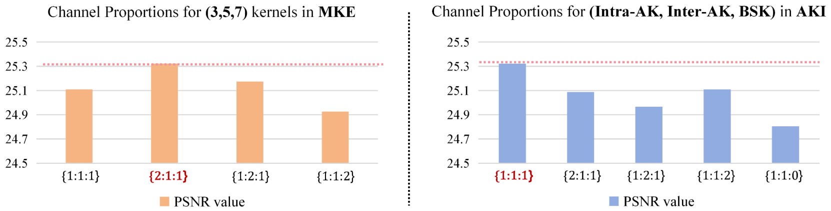

**Fig. 3.** PSNR under different channel ratios for MKE Block and AKI Block.

We also evaluate channel allocation strategies in MKE and AKI Blocks. As shown in Fig. 3, MKE performs optimally with a $2:1:1$ ratio, where dominant 3×3×3 kernels (50%) capture fine-grained features while 5×5×5 and 7×7×7 kernels (25% for each) efficiently model long-range dependencies, balancing multi-scale extraction ability and computational cost. In AKI Block, an equal $1:1:1$ allocation yields optimal results, enhancing the coordination of PET-specific self-calibration (via Intra-AK), PET-MRI cross-modal interaction (via Inter-AK), and stable feature preservation (via BSK). We also explored the “no BSK” variant (ratio $1:1:0$) and found it performs suboptimal, indicating that retaining BSK effectively mitigates over-reliance on AKs in early stages.

## 4 Conclusion

In this paper, we proposed a novel framework named MAK-GAN for asymmetric multi-modal PET reconstruction by leveraging the auxiliary MRI guidance. Specifically, we introduced a Multi-Kernel Extraction (MKE) block to efficiently extract modality-specific features at multiple scales, and an Adaptive Kernel Interaction (AKI) block to preserve PET-specific characteristics while adaptively incorporating MRI’s anatomical information for further refinement. Additionally, we incorporated PET-centric optimization strategies to enhance reconstruction quality.

**Acknowledgments.** This work is supported by National Natural Science Foundation of China (NSFC 62371325, U23A20295, 62131015, 82441023, 82394432), Sichuan Science and Technology Program 2025NSFJQ0050, Sichuan Science and Technology Program 2024ZDZX0018, and Key Lab of Internet Natural Language Processing of Sichuan Provincial Education Department (No. INLP202402).

**Disclosure of Interests.** The authors have no competing interests to declare.

## References

1. Schwenck J, Sonanini D, Cotton J M, et al.: Advances in PET imaging of cancer. Nature Reviews Cancer 23(7), 474-490 (2023)
2. Reader A J, Zaidi H. Advances in PET image reconstruction. PET clinics 2(2), 173-190 (2007)
3. Feng, Q., Liu, H.: Rethinking PET image reconstruction: ultra-low-dose, sinogram and deep learning. In: Martel, A.L., et al. (eds.) MICCAI 2020, pp. 783-792. Springer, Cham (2020)
4. Reader A. J., Corda G., Mehranian A., et al.: Deep learning for PET image reconstruction. IEEE Transactions on Radiation and Plasma Medical Sciences, 5(1), 1-25 (2020)
5. Xiang L., Qiao Y., et al.: Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. Neurocomputing 267, 406-416 (2017)
6. Wang, Y., Yu, B., Wang, L., et al.: 3D conditional generative adversarial networks for high-quality PET image estimation at low dose. Neuroimage 174, 550-562 (2018)
7. Gong K, et al.: PET image denoising based on denoising diffusion probabilistic model. European Journal of Nuclear Medicine and Molecular Imaging, 51(2), 358-368 (2024)
8. Cui J, Zeng X, Zeng P, et al.: MCAD: Multi-modal Conditioned Adversarial Diffusion Model for High-Quality PET Image Reconstruction. In: Linguraru, M.G., et al. (eds.) MICCAI 2024, pp. 467-477. Springer, Cham (2024)
9. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015)
10. Zeng X., Zeng P., Tang C., et al.: DBTrans: A Dual-Branch Vision Transformer for Multi-Modal Brain Tumor Segmentation. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 502-512, Springer, Cham (2023)

11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2020)
12. Luo, Y., Wang, Y., Zu, C., et al.: 3D Transformer-GAN for high-quality PET reconstruction. In: de Bruijne, M., et al. (eds.) MICCAI 2021, vol. 12906, pp. 276–285. Springer, Cham (2021)
13. Zeng, P., Zhou, L., Zu, C., et al.: 3D CVT-GAN: a 3D convolutional vision transformer-GAN for PET reconstruction. In: Wang, L., et al. (eds.) MICCAI 2022, vol. 13436, pp. 516–526. Springer, Cham (2022)
14. Yu B., Zhou L., Wang L., et al.: Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. IEEE transactions on medical imaging, 38(7), 1750-1762 (2019)
15. Cui J, Zeng P, Zeng X, et al.: Trido-former: A triple-domain transformer for direct pet reconstruction from low-dose sinograms. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 184-194, Springer, Cham (2023)
16. Cui J, Zeng P, Zeng X, et al.: Prior knowledge-guided triple-domain transformer-GAN for direct PET reconstruction from low-count sinograms. IEEE transactions on medical imaging, 43(12), 4174-4189 (2024)
17. Han K., Wang Y., Chen H., et al.: A survey on vision transformer. IEEE transactions on pattern analysis and machine intelligence 45(1), 87-110 (2022)
18. Zeng P, Zeng X, Wang Y, et al.: Multi-modal Long-Short Distance Attention-based Transformer-GAN for PET Reconstruction with Auxiliary MRI. IEEE Transactions on Circuits and Systems for Video Technology (2025)
19. Aiello M., Cavaliere C., Fiorenza D., et al.: Neuroinflammation in neurodegenerative diseases: current multi-modal imaging studies and future opportunities for hybrid PET/MRI. Neuroscience, 403, 125-135 (2019)
20. Wang, Y., Zhou, L., Yu, B., et al.: 3D auto-context-based locality adaptive multi-modality GANs for PET synthesis. IEEE transactions on medical imaging, 38(6), 1328-1339 (2018).
21. Wang Y., Luo Y., Zu C., et al.: 3D multi-modality Transformer-GAN for high-quality PET reconstruction. Medical Image Analysis 91, 102983 (2024)
22. Cui J., Wang Y., Zhou L., et al.: 3D Point-Based Multi-Modal Context Clusters GAN for Low-Dose PET Image Denoising. IEEE Transactions on Circuits and Systems for Video Technology, 34(10), 9400-9413 (2024)
23. Jiang C., Pan Y., Liu M., et al.: PET-diffusion: Unsupervised PET enhancement based on the latent diffusion model. In: Greenspan, H., et al. (eds.) MICCAI 2023, pp. 3-12. Springer, Cham (2023)
24. Gan W., Xie H., von Gall C., et al.: Pseudo-MRI-guided PET image reconstruction method based on a diffusion probabilistic model. IEEE Transactions on Radiation and Plasma Medical Sciences (2025)
25. Chen J., Wang X., Guo Z., et al.: Dynamic region-aware convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8064-8073 (2021)
26. Duan Z., Zhang T., Luo X., et al.: DCKN: Multi-focus image fusion via dynamic convolutional kernel network. Signal Processing 189, 108282 (2021)
27. Lei T., Zhang D., Du X., et al.: Semi-supervised medical image segmentation using adversarial consistency learning and dynamic convolution network. IEEE transactions on medical imaging 42(5), 1265-1277 (2022)
28. Huang J., Wang M., Ju H., et al.: SD-CNN: A static-dynamic convolutional neural network for functional brain networks. Medical Image Analysis 83, 102679 (2023)

29. Guan Y., Xu R., Yao M., et al.: Mutual-guided dynamic network for image fusion. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 1779-1788 (2023)
30. Gao W., Zhang X., Yang L., et al.: An improved Sobel edge detection. In: 3rd International conference on computer science and information technology, pp.67-71. IEEE (2010)