HKAFusion: Hybrid Kernel Attention Network for Medical Image Fusion

1st Siqi Zeng
University of Science and Technology Beijing
Beijing, China
M202410650@xs.ustb.edu.cn

2nd Hongjun Liu
University of Science and Technology Beijing
Beijing, China
D202210386@xs.ustb.edu.cn

3rd Chao Yao
University of Science and Technology Beijing
Beijing, China
yaochao@ustb.edu.cn

4th Jinsheng Sun
University of Science and Technology Beijing
Beijing, China
sunpro108@163.com

5th Sha Tao
University of Science and Technology Beijing
Beijing, China
uiqtuf@163.com

6th Xiaojuan Ban
University of Science and Technology Beijing
Beijing, China
banxj@ustb.edu.cn

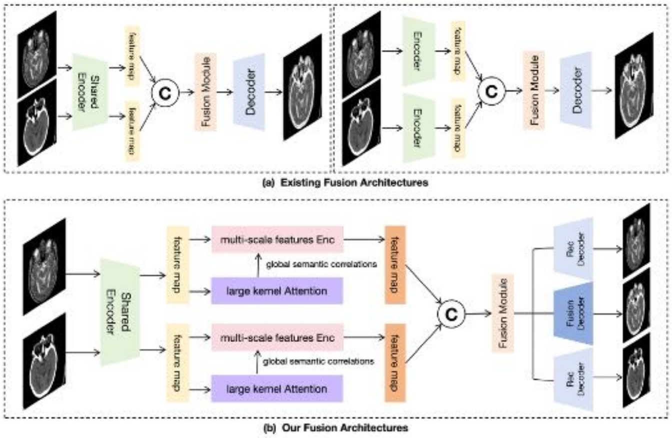
2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3315-1557-7/25/$31.00 ©2025 IEEE | DOI: 10.1109/BIBM66473.2025.11357018

***Abstract*—Multimodal medical image fusion is crucial for clinical diagnosis and treatment, as it integrates complementary information from multiple imaging modalities into a single image. However, existing fusion methods still struggle to preserve both fine lesion details and global anatomical structures across modalities. They often over-smooth small structures, blur organ boundaries, and introduce inconsistencies between structural and functional information. To address this problem, we introduce a hybrid kernel attention mechanism that couples small-kernel convolutions with large-kernel attention in a unified block, rather than relying on a fixed receptive field. It leverages scale-aware relationships between lesion-level textures, organ boundaries, and global anatomy across modalities, embedding these relationships into the fused representation. Concretely, A U-shaped encoder–decoder backbone is adopted with Restormer-based shallow feature extraction and several scale-aware hybrid kernel attention blocks are stacked at multiple resolutions. The resulting modality-specific features are then fed into a multimodal fusion module, which adaptively couples them via cross-modal interaction and reconstruction branches to enforce structural and semantic consistency.. We adopt a comprehensive evaluation protocol spanning eight widely used fusion metrics on MRI–CT, MRI–PET, and MRI–SPECT benchmarks. Our proposed method achieves consistent improvements over recent state-of-the-art methods, yielding up to 41.9% relative gain in mutual information compared with CDDFusion on the MRI–PET dataset, demonstrating its superiority for multimodal medical image fusion. Moreover, qualitative visual comparisons show sharper anatomical boundaries and more distinguishable lesion regions, indicating that HKAFusion preserves clinically relevant details more faithfully.**

***Index Terms*—medical image fusion, multiscale convolution, Attention module, deep neural network, Feature extraction**

I. Introduction

Medical image processing serves as a cornerstone technology in modern healthcare systems, with widespread applications in image enhancement [34], segmentation [2], [36], [37], registration [3], and computer-aided diagnosis [4]. Benefiting from the development of various imaging modalities such as CT, MRI, and PET, different modalities provide complementary information from multiple perspectives [5], including anatomical structure, tissue density, and functional metabolism, thereby greatly expanding the potential for accurate lesion identification and assessment. To achieve a unified representation of the advantages offered by multiple modalities, medical image fusion [6] has emerged as a crucial technique. Its objective is to synergistically integrate the key information from heterogeneous modalities into a single image, thereby offering a more discriminative and informative visual basis for clinical decision-making. [7], [43]



(a) Existing Fusion Architectures

(b) Our Fusion Architectures

Fig. 1. Workflow comparison of our proposed HKAFusion with existing image fusion approaches.

Unlike conventional image processing tasks, medical image fusion requires not only the preservation of fine-grained structural details but also the dynamic integration of cross-modal semantic representations to ensure anatomical coherence. The core challenge lies in achieving multi-scale feature modeling that simultaneously captures local textures and global dependencies, while maintaining semantic alignment between modalities. Traditional approaches often treat these objectives independently, leading to fused images with structural distortions or semantic inconsistencies in critical regions.

Current fusion methods predominantly rely on single-paradigm feature extractors: While CNNs [17], [31]–[33] excel at local detail preservation, their limited receptive fields hinder cross-region relationship modeling. Conversely, Transformers [9], [10] leverage long-range dependencies but often sacrifice edge precision and spatial continuity. Hybrid CNN-Transformer architectures [11], [40], [41] attempt to mitigate these limitations, yet they fail to adaptively balance modality-specific features during fusion, particularly in handling heterogeneous semantic distributions (e.g., functional vs. structural modalities). This gap underscores the need for a unified attention mechanism capable of multi-scale dynamic fusion while enforcing cross-modal semantic consistency.

Moreover, the fidelity of fused medical images hinges not on simple intensity superposition, but on the hierarchical integration of complementary multimodal features—where structural continuity and semantic discriminability must coexist. Existing methods typically follow a pipeline where features from each modality are extracted separately and then simply concatenated for decoding, often treating modalities as isolated streams, leading to spatially inconsistent structural overlaps (e.g., misaligned organ boundaries in MRI-CT fusion) or semantically ambiguous feature blending (e.g., metabolic-active regions in PET losing anatomical specificity). To address this, a paradigm shift is imperative: A fusion framework must dynamically coordinate structural granularity and semantic granularity, which is exactly the gap our hybrid kernel attention mechanism aims to bridge—enabling content-aware feature reinforcement and cross-modal redundancy suppression through learnable multi-scale interactions, as shown in the Fig. 1.

In summary, the fundamental objective of medical image fusion lies in achieving dual optimization of fine-grained detail fidelity and high-level semantic complementarity, thereby enhancing the clinical interpretability and decision-support capability of fused images. This process presents two critical challenges: (1) how to accurately identify and fuse semantically critical regions from heterogeneous modalities to achieve complementary representation of deep semantic features; (2) how to dynamically preserve cross-modal structural details and boundary coherence during fusion to prevent spatial degradation and texture loss. Conventional fusion approaches relying on fixed-scale convolutions or single-mechanism attention frameworks struggle to simultaneously model multi-scale contextual relationships while ensuring inter-modal structural alignment, frequently resulting in blurred edges, incomplete semantic representation, or information conflicts in fused outputs. Although Transformer-based models [12] demonstrate superior global modeling capabilities, they face inherent limitations in balancing multi-scale representation with computational efficiency - particularly when processing high-resolution medical images with structurally ambiguous regions.

To address the aforementioned challenges, we propose a hybrid-kernel attention framework for medical image fusion. The method leverages the synergistic benefits of multi-scale convolutional kernels and attention mechanisms, enabling deep integration of fine-grained features and semantic structures across different modalities. Specifically, we design a U-shaped encoder–decoder architecture capable of adaptively extracting and aggregating complementary information from multiple modalities.Building on this architecture, we introduce a novel Scale aware Kernel Attention Block. This block employs parallel small and large convolutional kernels to generate attention maps that focus on local details and global semantics, respectively. The small kernels are tailored to capture high frequency textures and edge information, whereas the large kernels excel at modeling long-range dependencies and holistic anatomical structures. These multi-scale features are subsequently fused through an attention gating mechanism, where the adaptive weighting strategy intelligently balances the contributions of different kernel branches based on local regional characteristics.Extensive experiments demonstrate that our model consistently outperforms several state-of-the-art methods across various medical image fusion tasks. The main contributions of this work are summarized as follows:

- We develop a hybrid kernel attention-based image fusion framework, termed HKAFusion, which achieves more precise fusion of multimodal medical images.
- We design a Scale-aware Kernel Attention Block that adaptively balances fine-grained details and high-level semantic features through a multi-scale kernel attention mechanism.
- We embed the SKA Block into a U-shaped architecture by stacking multiple SKA blocks at different resolutions, which enables effective integration of complementary information across modalities and improves the representational capacity of the fused images.

II. Related Works

In this section, we first review deep learning-based image fusion methods. Then, we introduce the large-kernel convolutional attention mechanism.

A. Deep Learning-based Image Fusion Method

Current medical image fusion methods are primarily based on CNNs [17], [31]–[33], Transformers [9], [10], GANs [14], [35], diffusion models [15], [38], [39], and their hybrid architectures [11], [40]. IFCNN [17] proposed a fully convolutional framework that employs a pretrained ResNet for feature extraction and applies pixel-level fusion rules to integrate multi-source image information. This significantly reduces artifacts and enhances detail preservation, but the limited receptive field of fully convolutional methods hinders their ability to capture global structural information. SwinFusion [16] leverages the shifted window mechanism of Swin Transformer [9] to construct a cross-domain long-range learning module, combining intra-domain self-attention and inter-domain cross-attention. It significantly improves global feature interaction capabilities but exhibits weak perception of local details and high computational complexity. TADAL [14] enhances object-aware fusion by imposing stronger adversarial constraints on salient regions, ensuring clear targets and natural background textures. Dif-Fusion [15] introduces a novel diffusion-based method to model the distribution of multi-channel input data, enhancing multi-source information aggregation and color fidelity. Although GAN-based and diffusion-based approaches produce visually appealing results, they often focus on overall image realism while lacking precise modeling of critical medical details and semantic regions. CDDFuse [11] combines CNNs to extract low-level local features from multimodal images and Transformers to capture long-range dependencies across modalities, addressing challenges in multimodal feature alignment and retention. However, such hybrid models are often structurally complex, demand high computational resources and large-scale training data, and still face challenges in coordinating detail and global information across scales and maintaining training stability.

B. Large-kernel Convolutional Attention Mechanism

In recent years, studies [6], [18], [19], [42] have shown that enlarging convolutional kernel sizes can significantly enhance the performance of visual models. This approach can partially mimic the large-range spatial modeling capabilities of Transformers, improving long-range dependency capture while offering superior computational efficiency and lower model complexity compared to traditional Transformer architectures. As a result, many works [20]–[22] have focused on expanding the receptive field of convolutional neural networks (CNNs) by increasing kernel sizes. For instance, RepLKNet [20] employs ultra-large kernels (e.g., 31×31) and achieves performance comparable to Vision Transformers (ViTs). To mitigate the computational burden of large kernels, the Large Kernel Attention (LKA) [21] mechanism was proposed, which maintains a large receptive field while significantly reducing computational complexity. However, the depthwise convolution layers in LKA lead to quadratic growth in computation and memory usage as the kernel size increases. To address this, LSKA [22] improves efficiency by decomposing the 2D depthwise convolution kernel into cascaded horizontal and vertical 1D kernels. Nevertheless, LSKA is designed for a single-scale receptive field, making it insufficient for capturing the multi-scale features that are critical in medical imaging.

III. Method

In this section, we present HKAFusion, a novel multimodal medical image fusion framework. We first provide an overview of the overall network architecture. Then, we describe the design of the SKA block and the feed-forward module [21] in detail. Finally, we introduce the implementation of the multimodal feature fusion module.

A. Overview of the Framework

Our network architecture consists of two main modules: a multi-modal feature extraction module and a feature fusion module, as shown in Fig. 2.

During the feature extraction stage, two Restormer [10] blocks are first employed to extract shallow features from the input multi-modal medical images. Restormer uses cross-channel self-attention to model global contextual information. It is capable of extracting semantically rich shallow features from high-resolution images while maintaining high computational efficiency. Subsequently, the network enters four Scale-aware Kernel Attention stages, performing feature extraction on the two modality images at four different resolutions: $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$, where H and W denote the height and width of the input image, respectively. Each stage starts with a downsampling operation to reduce spatial resolution by adjusting the stride. The number of channels increases progressively as the resolution decreases. Then, several SKA blocks are stacked to model both local and long-range dependencies. Inside each SKA block, the spatial size and number of channels remain the same, helping to improve the semantic representation.

The feature fusion stage consists of five steps, each comprising a feature fusion module and an upsampling operation. The module first performs multi-level fusion of multimodal features by leveraging the corresponding resolution features obtained from the encoder, and then generates the final fused image through the decoder. Furthermore, to ensure information integrity during feature extraction and fusion, we have also designed an image reconstruction branch, which employs a reconstruction decoder to recover the original input images.

B. Scale-aware Kernel Attention Block

The core design of the SKA block lies in the synergistic integration of the multi-scale small convolutional kernels and the large-kernel attention mechanism within the SKA module. As shown in Fig. 2, each block maintains a consistent spatial resolution and channel count throughout, and comprises a sequential stack of: Batch Normalization, a 1×1 convolution, GELU activation, the Scale-aware Kernel Attention Module, and a Feed-Forward Network for feature extraction.

Scale-aware Kernel Attention Module Our module employs a multi-scale strategy to enhance the integration of large-kernel convolutional attention and small-kernel convolutions, with specialized optimization for medical image fusion. To address the significant variations in anatomical structures and feature scales in medical images, the module leverages convolution kernels of different sizes to simultaneously capture multi-scale local features and global attention maps across modalities. This design enables the extraction of fine-grained local details while modeling long-range structural dependencies.

Specifically, the module incorporates parallel convolutions with kernel sizes such as 3×3, 5×5, 7×7, each responsible for extracting local features under different receptive fields. Since large convolution kernels can significantly enlarge the receptive field and more effectively capture long-range dependencies and global structural information across organs and regions, we introduce a depthwise separable large kernel convolution to generate an attention map that provides global

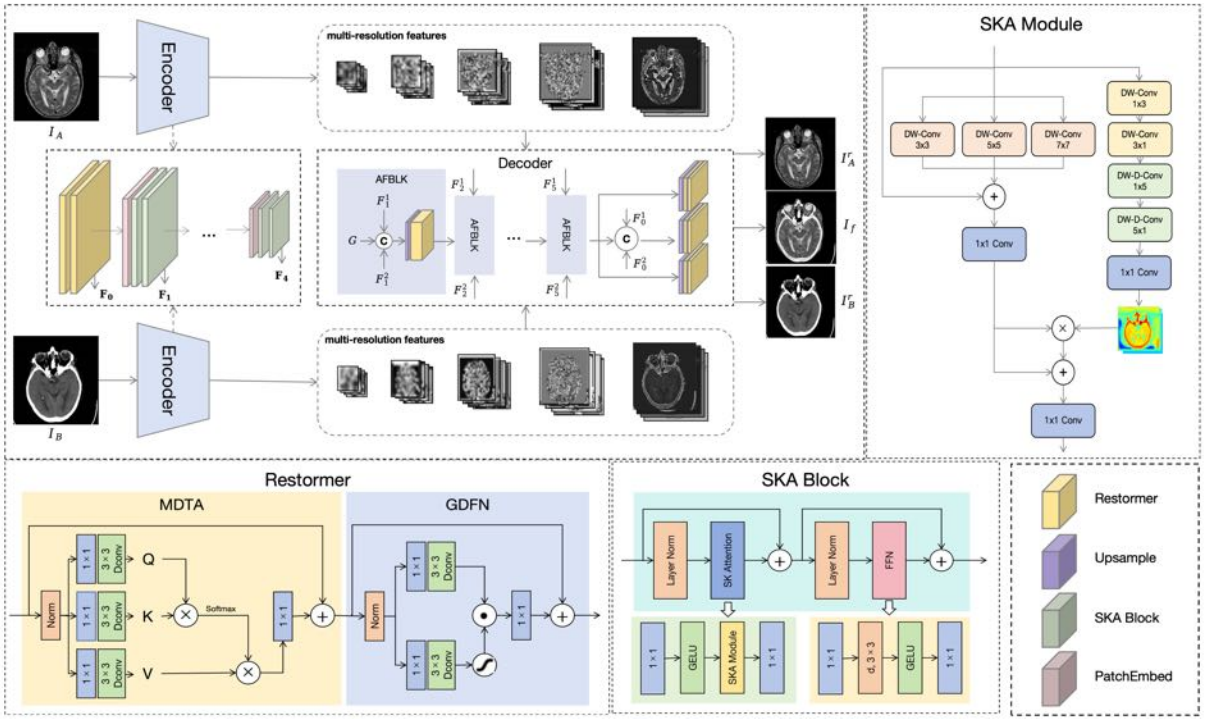


Fig. 2. Overall framework of the proposed method.

guidance. By adaptively fusing and weighting this attention map with the multi-scale local features, the module preserves fine textures from the source images while establishing global semantic correlations, thus producing richer and more discriminative feature representations for fusion. The SKA module in the $n-th$ SKA block of the $l-th$ stage can be mathematically expressed as follows:

$$x_{\mathrm{d}} = \mathrm{Conv}_{1\times1}\left(\mathbf{x} + \sum_{m=1}^{3} \mathrm{DWConv}_{k^{(m)}\times k^{(m)}}(\mathbf{x})\right) \tag{1}$$

$$\mathrm{Attn} = \mathrm{Conv}_{\mathrm{DW}}^{(2d-1)\times1}\big(\mathrm{Conv}_{\mathrm{DW}}^{1\times(2d-1)}(\mathbf{x})\big) \tag{2}$$

$$\mathrm{Attention} = \mathrm{Conv}_{\text{DW-D}}^{\lfloor\frac{k_i}{d}\rfloor\times1}\big(\mathrm{Conv}_{\text{DW-D}}^{1\times\lfloor\frac{k_i}{d}\rfloor}(\mathrm{Attn})\big) \tag{3}$$

$$x_{\mathrm{out}} = \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}\left(Attention \otimes x_{\mathrm{d}}, x_{\mathrm{d}}\right)\right) \tag{4}$$

here, $x \in \mathbb{R}^{C\times H\times W}$ is the input feature, x_{d} denotes the multi-scale local features, $Attention \in \mathbb{R}^{C\times H\times W}$ is the attention map, with each value representing the relative importance of the corresponding feature. The operator $\otimes$ denotes the element-wise multiplication operation.

Feed-Forward Network. To enhance the feature processing capability, we introduce the FFN, as shown in Fig. 2, which is a guided feed-forward network. The primary function of the FFN module is to apply further non-linear transformations to the features processed by the attention mechanism, thereby improving the feature representation and the model's learning capacity. Specifically, the FFN operates by processing the input features layer by layer through multiple fully connected layers and non-linear activation functions (such as ReLU), enabling the model to capture more complex patterns and higher-order features. Given an input feature map X , the FFN processes it as follows:

$$FFN = Conv_{1\times1}\left(GeLU\left(Conv_{3\times3,d}\left(Conv_{1\times1}(x)\right)\right)\right) \tag{5}$$

here, $Conv_{3\times3,d}$ represents a dilated convolution with a kernel size of 3 and a dilation rate of 2. The FFN is typically employed as a key component for feature extraction and transformation. By integrating the FFN with the SKA, the network's performance can be further enhanced.

C. Feature Fusion Module

As shown in Fig. 2, we feed features F_i^A and F_i^B into the FFM module for multimodal feature fusion. The FFM comprises five AFBLK modules and three decoders. For the $i-th$ AFBLK, its inputs include F_i^A and F_i^B, and the output G_{i-1} from the previous block, producing the output G_i:

$$G_i = R_e\left(\mathrm{Concat}\left(F_{i-1}^A, F_{i-1}^B, \uparrow_{\times2}(G_{i-1})\right)\right) \tag{6}$$

where R_e denotes the encoder composed of Restormer blocks. The index i starts from 2, G_1 is a zero matrix when $i = 2$. The output of the final AFBLK modules is denoted as G_5. Subsequently, we concatenate G_5 with F_0^A and F_0^B, and feed the concatenated result into a fusion image generator composed of a Restormer block, a convolutional layer, and a sigmoid activation function to produce the fused image I_f.

Furthermore, to ensure that no information is lost during feature extraction and fusion reconstruction, we introduce two additional reconstruction decoders. These decoders share an identical architecture with the fusion image generator and are responsible for reconstructing the two source images by taking the fused feature G_5 as their input.

D. Loss Function

The loss function plays a vital role in determining the performance of the fusion model. To regulate the feature transfer within the model, we have designed a composite loss function. We utilize the structural loss L_s to maximize the structural consistency between the fused image and the source images:

$$\mathcal{L}_s = \mathcal{L}_{ssim}\left(\boldsymbol{I}_f, \boldsymbol{I}_A\right) + \mu\,\mathcal{L}_{ssim}\left(\boldsymbol{I}_f, \boldsymbol{I}_B\right) \tag{7}$$

where I_f denotes the fused image, and I_A , I_B represent the two source modality images, respectively. The parameter μ is a hyperparameter used to balance the contributions of the two SSIM losses in L_S. To ensure that the fused image preserves clear edge details and good contrast, we introduce both gradient loss and pixel intensity loss:

$$\mathcal{L}_{inten} = \left\|\boldsymbol{I}_f - \max(\boldsymbol{I}_A, \boldsymbol{I}_B)\right\|_1 \tag{8}$$

$$\mathcal{L}_{grad} = \left\|\nabla\boldsymbol{I}_f - \max(\nabla\boldsymbol{I}_A, \nabla\boldsymbol{I}_B)\right\|_1 \tag{9}$$

Furthermore, to ensure that the comprehensive features of the source image learned during the self-reconstruction process are preserved, we introduce the reconstruction loss:

$$\mathcal{L}_1^r = \|I_A - I_A^{pre}\|_1 + 1 - \mathcal{L}_{ssim}\left(I_A, I_A^{pre}\right) \tag{10}$$

$$\mathcal{L}_2^r = \|I_B - I_B^{pre}\|_1 + 1 - \mathcal{L}_{ssim}\left(I_B, I_B^{pre}\right) \tag{11}$$

Therefore, the total loss of our method is as follows:

$$\text{loss} = \lambda_1 L_s + \lambda_2 L_{\text{grad}} + \lambda_3 L_{\text{inten}} + \mathcal{L}_1^r + \mathcal{L}_2^r \tag{12}$$

here, λ_k (where $k = 1, 2, 3$) represents the hyperparameters used to adjust the balance between the different loss functions.

IV. Experimental Results And Analysis

In this section, we conduct both qualitative and quantitative analyses to compare the performance of our proposed HKAFusion method with several state-of-the-art techniques on publicly available CT-MRI, PET-MRI, and SPECT-MRI datasets.

A. Experimental Setup

Atlas Dataset: In this study, we utilize the publicly available multimodal medical image dataset from Harvard, known as the Atlas dataset, which includes CT-MRI, PET-MRI, and SPECT-MRI modality pairs. The training set contains 185, 270, and 358 image pairs for CT-MRI, PET-MRI, and SPECT-MRI, respectively, while the corresponding test sets include 20, 43, and 74 pairs. All image pairs are resized to a resolution of 256×256. To enhance data diversity and improve model generalization, random rotations and flips are applied during each training epoch as data augmentation strategies. For RGB images, we convert them to the YCbCr color space and use only the Y (luminance) channel for training. PET and SPECT images are processed as grayscale images during training.

Implementation Details: During the training process, We adopt an end-to-end training strategy for each dataset, with 2000 epochs and a batch size of 8. The Adam optimizer is employed to update the model parameters, starting with an initial learning rate of 1×10^{-5}, decreasing to 1×10^{-7} over time. A cosine annealing learning rate scheduler is used to gradually reduce the learning rate over time, ensuring better training stability and convergence. In the loss function, we set three fixed hyperparameters: λ_1 = 1, λ_2 = 1, and λ_3 = 0.5. Additionally, a dynamic weighting factor μ is introduced to balance the structural similarity between the fused image and the source images. The value of μ is updated at the end of each epoch based on the average $SSIM$ between the fused image and the two source images, calculated as: $\mu = \frac{\sum_{n=1}^{N} \mathcal{L}_{\text{ssim}}^{(n)}(I_f, I_B)}{\sum_{n=1}^{N} \mathcal{L}_{\text{ssim}}^{(n)}(I_f, I_A)}$, where N denotes the number of training samples in each epoch. The proposed method is implemented using the PyTorch framework and trained on a single NVIDIA GeForce RTX 4090 GPU.

Evaluation Metrics: To quantitatively evaluate the fusion performance of HKAFusion, we adopt eight commonly used image quality metrics: Mutual Information (MI), Spatial Frequency (SF), Visual Information Fidelity (VIF), Average Gradient (AG), Gradient-based Fusion Performance ($Q_{AB/F}$), Chen-Varshney Metric (Q_{CV}), Correlation Coefficient (CC), and Structural Similarity Index Measure ($SSIM$). Except for Q_{CV}, higher values of these metrics generally indicate better fusion quality.

B. Fusion Results

In this section, we conduct both qualitative and quantitative comparisons with six state-of-the-art methods, including U2Fusion [23], DDFM [24], ReCoNet [25], CDDFuse [11] and MsgFusion [26]. The parameters for each method are strictly set according to the corresponding literature.

a) Qualitative comparison: Fig.3 and Fig.4 show the qualitative comparison results of HKAFusion on MRI-CT and MRI-PET/SPECT fusion tasks. To support analysis, we highlight important anatomical regions in the source and fused images using red boxes, including brain gray matter and tumor areas. The results indicate that U2Fusion [23] and DDFM [24] tend to blur edge structures and weaken details, making it difficult to clearly represent lesion boundaries. ReCoNet [25] and MsgFusion [26] can retain structural information from both modalities, but the fused images often show low local contrast and lack rich texture. CDDFuse [11] achieve better visual perception but may introduce redundant information or weaken target expression at the semantic level. Compared with these methods, HKAFusion preserves the high-density bone structures from CT and the soft tissue details from MRI more effectively in the MRI-CT fusion. In the MRI-PET fusion, it enhances the expression of metabolic hotspots from PET while maintaining the structural information from MRI. This demonstrates stronger consistency between structure and semantics and better ability to highlight discriminative regions.

b) Quantitative comparison: The experimental results in Table.I show that the proposed method achieves the best

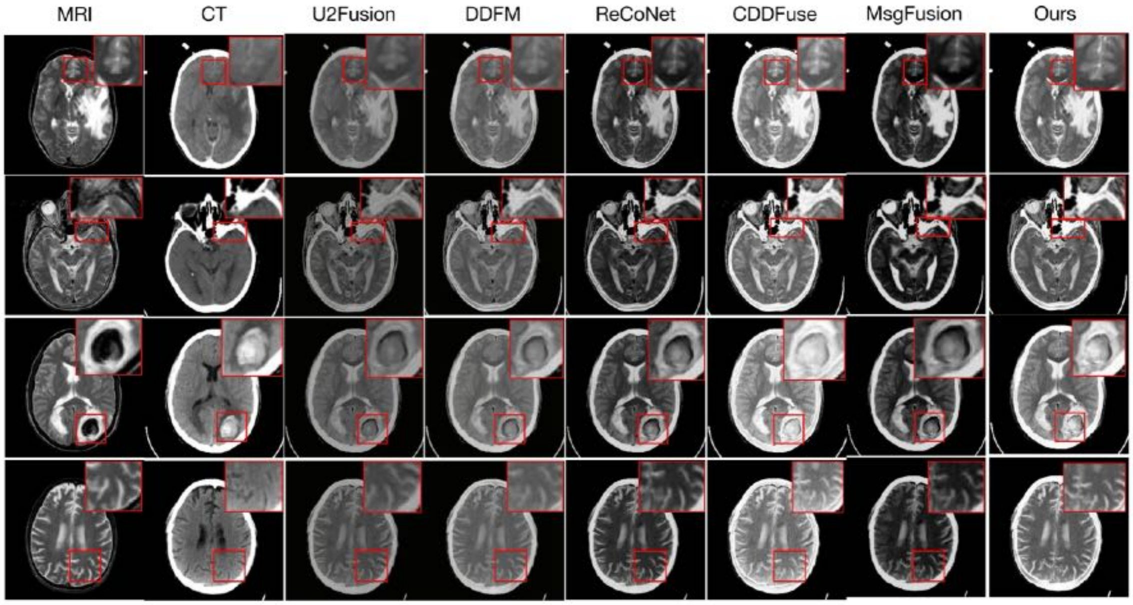


Fig. 3. Qualitative comparisons of our HKAFusion with five existing methods on four representative CT-MRI image pairs.

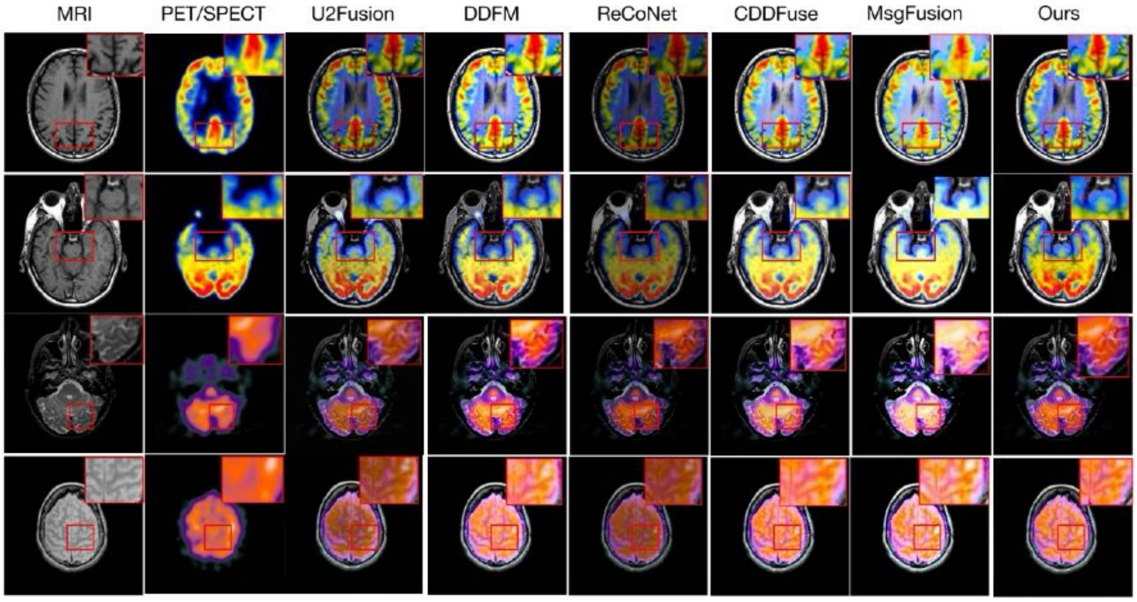


Fig. 4. Qualitative comparisons of our HKAFusion with five existing methods on four representative PET/SPECT-MRI image pairs.

performance on most metrics. In particular, it performs well on VIF and MI, reaching 0.81 / 0.83 and 2.85 / 2.88 on the MRI-CT and MRI-PET tasks, respectively. These results indicate that the method can better preserve structural information and effectively fuse multi-source data. In addition, the SF and Q_{CV} scores are also higher than other methods, showing that the method improves detail clarity and texture representation. It can also better retain and enhance complementary features between modalities, supporting the joint modeling of semantic and structural information.

C. Ablation Study

To evaluate the effectiveness of the key components in our proposed model, we conducted a series of systematic ablation studies on the MRI-CT dataset. As shown in Table II, these experiments focus on two main aspects: the design of the hybrid kernel attention mechanism and the construction of the multimodal feature fusion module.

We first compare three different attention configurations, completely removing the attention module, replacing it with Transformer-based self-attention, and employing our proposed

TABLE I
AVERAGE SCORES OF DIFFERENT MODELS ON MEDICAL IMAGE FUSION. RED INDICATES THE BEST RESULT. BLUE INDICATES THE SUBOPTIMAL RESULT.

Dataset: MRI-CT Medical Image Fusion								
	$MI\uparrow$	$SF\uparrow$	$VIF\uparrow$	$AG\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$CC\uparrow$	$Q_{SSIM}\uparrow$
U2Fusion [23]	2.08	22.54	0.75	6.63	0.46	1352.34	0.52	0.49
DDFM [24]	2.56	21.15	0.78	6.50	0.51	1089.20	0.70	0.91
ReCoNet [25]	2.03	20.16	0.40	6.35	0.42	2648.20	0.50	1.29
CDDFusion [11]	2.61	33.83	0.61	7.33	0.58	976.39	0.76	1.34
MsgFusion [26]	2.38	20.43	0.59	6.44	0.41	2090.40	0.63	0.39
Ours	2.85	34.83	0.81	7.43	0.63	890.00	0.69	1.54
Dataset: MRI-PET Medical Image Fusion								
	$MI\uparrow$	$SF\uparrow$	$VIF\uparrow$	$AG\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$CC\uparrow$	$Q_{SSIM}\uparrow$
U2Fusion [23]	1.69	23.27	0.40	8.70	0.49	1023.50	0.44	1.39
DDFM [24]	1.84	22.65	0.52	8.63	0.57	879.20	0.59	1.43
ReCoNet [25]	1.51	21.72	0.44	8.61	0.51	549.26	0.43	1.40
CDDFusion [11]	2.03	29.57	0.71	8.74	0.71	360.23	0.67	1.49
MsgFusion [26]	1.74	22.53	0.76	8.53	0.45	486.40	0.48	1.04
Ours	2.88	34.64	0.83	8.96	0.67	361.85	0.77	1.36
Dataset: MRI-SPECT Medical Image Fusion								
	$MI\uparrow$	$SF\uparrow$	$VIF\uparrow$	$AG\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$CC\uparrow$	$Q_{SSIM}\uparrow$
U2Fusion [23]	1.68	19.58	0.48	5.20	0.57	328.43	0.65	1.41
DDFM [24]	1.95	18.52	0.54	5.14	0.62	230.43	0.77	1.45
ReCoNet [25]	1.50	17.40	0.46	5.12	0.54	579.26	0.63	1.40
CDDFusion [11]	2.49	20.87	0.97	5.38	0.78	120.38	0.88	1.48
MsgFusion [26]	1.83	18.26	0.89	5.12	0.53	490.60	0.69	1.12
Ours	2.83	22.72	0.92	5.50	0.72	106.09	0.85	1.50

TABLE II
AVERAGE SCORES OF THE ABLATION EXPERIMENTS.

	$MI\uparrow$	$SF\uparrow$	$VIF\uparrow$	$AG\uparrow$	$Q_{AB/F}\uparrow$	$Q_{CV}\downarrow$	$CC\uparrow$	$Q_{SSIM}\uparrow$
w/o attention mechanism	2.14	23.52	0.55	6.92	0.46	1843.95	0.56	1.14
Self-attention mechanism	2.76	31.17	0.78	7.41	0.53	1049.36	0.70	1.36
Direct Concatenation	2.37	24.49	0.52	7.39	0.48	1237.30	0.51	1.38
w/o reconstruction branch	2.72	34.69	0.80	7.32	0.58	917.00	0.64	1.52
(3, 3, 3, 7)	2.70	30.96	0.77	7.39	0.42	897.45	0.69	1.29
(3, 5, 7, 7))	2.71	33.83	0.79	7.38	0.68	920.65	0.74	1.34
(3, 5, 7, 11)	2.85	34.83	0.81	7.43	0.63	890.00	0.69	1.54

hybrid-kernel attention mechanism. The results demonstrate that our attention design provides a clear advantage in fusion performance.

To further investigate the impact of kernel size on fusion quality, several model variants are contructed that differ only in the convolutional kernel size. The experimental results indicate that incorporating multi-scale local feature extraction branches significantly improves the quality of the fused images. Moreover, using a larger kernel (e.g. 11×11) yields slightly better fusion performance compared with a 7×7 kernel, confirming the importance of a larger receptive field for capturing cross-modality contextual information. In addition, to assess the effectiveness of the FFM, we remove the FFM and perform fusion using simple feature concatenation. The experimental results in the table II show a clear performance degradation, indicating that the FFM is more effective than direct concatenation in integrating complementary information from different modalities.

To evaluate the contribution of the reconstruction branch, we performed an ablation experiment by removing this branch and assessing its impact on feature extraction and fusion quality. Without reconstruction supervision, the fused images exhibit noticeable degradation in structural details and texture consistency. This is because the reconstruction branch provides an additional constraint during training, encouraging the network to preserve information throughout the feature extraction and fusion process and preventing the loss of critical details in deep feature representations.

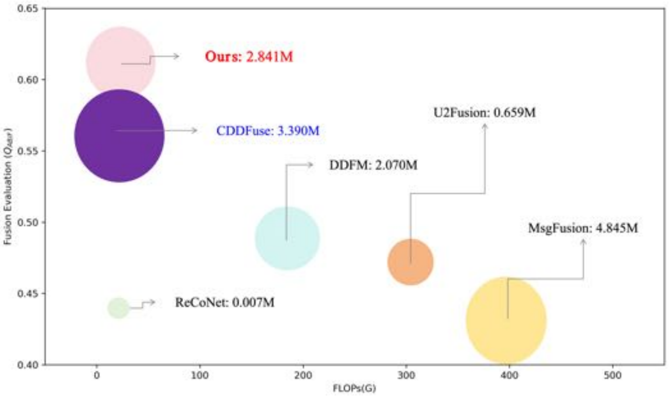


Fig. 5. Model complexity analysis.

D. Analysis of Model Complexity

In addition to fusion performance, the number of parameters and computational complexity are critical factors for practical applications. To further evaluate the trade-off between efficiency and effectiveness, we adopt the $Q_{AB/F}$ metric as the evaluation criterion and compare several state-of-the-art fusion models in terms of parameter scale and computational cost, aiming to comprehensively assess their fusion capability and suitability for real-world deployment.

As illustrated in Fig.5, although HKAFusion incorporates large-kernel attention and multi-scale design, it maintains a compact model size and low computational cost thanks to its efficient modular architecture and optimized convolutional operations. Experimental results demonstrate that HKAFusion achieves high fusion performance while significantly reducing parameter count and computational complexity.

V. CONCLUSION

This paper proposes a hybrid kernel attention-based medical image fusion framework, named HKAFusion, which aims to simultaneously enhance structural detail representation and semantic alignment across multimodal images. We design a SKA block, which integrates both large and small kernel attention mechanisms to adaptively capture fine-grained details and long-range dependencies. Experimental results demonstrate that the proposed method achieves more stable performance in terms of edge clarity, texture reconstruction, and lesion information preservation, showing promising potential for clinical applications.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant Nos. U22A2022, 62406024), and was conducted in part at the MOE Key Laboratory of Advanced Materials and Devices for Post-Moore Chips, the Beijing Key Laboratory of Big Data Innovation and Application for Skeletal Health Medical Care, and the Beijing Advanced Innovation Center for Materials Genome Engineering.

References

[1] Wang D, Liu J, Fan X, et al. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration[J]. arXiv preprint arXiv:2205.11876, 2022.

[2] Wang R, Lei T, Cui R, et al. Medical image segmentation using deep learning: A survey[J]. IET image processing, 2022, 16(5): 1243-1267.

[3] Chen J, Frey E C, He Y, et al. Transmorph: Transformer for unsupervised medical image registration[J]. Medical image analysis, 2022, 82: 102615.

[4] Ker J, Wang L, Rao J, et al. Deep learning applications in medical image analysis[J]. Ieee Access, 2017, 6: 9375-9389.

[5] Meher B, Agrawal S, Panda R, et al. A survey on region based image fusion methods[J]. Information Fusion, 2019, 48: 119-132.

[6] Ranftl R, Bochkovskiy A, Koltun V. Vision transformers for dense prediction[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 12179-12188.

[7] Azam M A, Khan K B, Salahuddin S, et al. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics[J]. Computers in biology and medicine, 2022, 144: 105253.

[8] Tang, W.; He, F.; Liu, Y.; and Duan, Y. 2022b. MATR: Multimodal medical image fusion via multiscale adaptive transformer. IEEE Transactions on Image Processing, 31: 5134– 5149.

[9] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[10] Zamir S W, Arora A, Khan S, et al. Restormer: Efficient transformer for high-resolution image restoration[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5728-5739.

[11] Zhao Z, Bai H, Zhang J, et al. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 5906-5916.

[12] Parmar N, Vaswani A, Uszkoreit J, et al. Image transformer[C]//International conference on machine learning. PMLR, 2018: 4055-4064.

[13] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[14] Liu J, Fan X, Huang Z, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 5802-5811.

[15] Yue J, Fang L, Xia S, et al. Dif-fusion: Toward high color fidelity in infrared and visible image fusion with diffusion models[J]. IEEE Transactions on Image Processing, 2023, 32: 5705-5720.

[16] Ma J, Tang L, Fan F, et al. SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

[17] Zhang Y, Liu Y, Sun P, et al. IFCNN: A general image fusion framework based on convolutional neural network[J]. Information Fusion, 2020, 54: 99-118.

[18] Yan H, Li Z, Li W, et al. ConTNet: Why not use convolution and transformer at the same time?[J]. arxiv preprint arxiv:2104.13497, 2021.

[19] Luo W, Li Y, Urtasun R, et al. Understanding the effective receptive field in deep convolutional neural networks[J]. Advances in neural information processing systems, 2016, 29.

[20] Ding X, Zhang X, Han J, et al. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11963-11975.

[21] Guo M H, Lu C Z, Liu Z N, et al. Visual attention network[J]. Computational visual media, 2023, 9(4): 733-752.

[22] Lau K W, Po L M, Rehman Y A U. Large separable kernel attention: Rethinking the large kernel attention design in cnn[J]. Expert Systems with Applications, 2024, 236: 121352.

[23] Xu H, Ma J, Jiang J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 44(1): 502-518.

[24] Zhao Z, Bai H, Zhu Y, et al. DDFM: denoising diffusion model for multi-modality image fusion[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 8082-8093.

[25] Huang Z, Liu J, Fan X, et al. Reconet: Recurrent correction network for fast and efficient multi-modality image fusion[C]//European conference on computer Vision. Cham: Springer Nature Switzerland, 2022: 539-555.

[26] Wen J, Qin F, Du J, et al. MsgFusion: Medical semantic guided two-branch network for multimodal brain image fusion[J]. IEEE Transactions on Multimedia, 2023, 26: 944-957.

[27] Xu H, Yuan J, Ma J. Murf: Mutually reinforcing multi-modal image registration and fusion[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(10): 12148-12166.

[28] Tang L, Deng Y, Ma Y, et al. SuperFusion: A versatile image registration and fusion network with semantic awareness[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(12): 2121-2137.

[29] Tang W, He F, Liu Y, et al. MATR: Multimodal medical image fusion via multiscale adaptive transformer[J]. IEEE Transactions on Image Processing, 2022, 31: 5134-5149.

[30] Wang D, Liu J, Ma L, et al. Improving misaligned multi-modality image fusion with one-stage progressive dense registration[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 10944-10958.

[31] Li H, Wu X J, Kittler J. RFN-Nest: An end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86.

[32] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[33] Liang P, Jiang J, Liu X, et al. Fusion from decomposition: A self-supervised decomposition approach for image fusion[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 719-735.

[34] Ma Y, Liu J, Liu Y, et al. Structure and illumination constrained GAN for medical image enhancement[J]. IEEE Transactions on Medical Imaging, 2021, 40(12): 3955-3967.

[35] Mao X, Li Q, Xie H, et al. Least squares generative adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2794-2802.

[36] Chen J, Mei J, Li X, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers[J]. Medical Image Analysis, 2024, 97: 103280.

[37] Rahman M M, Munir M, Marculescu R. Emcad: Efficient multi-scale convolutional attention decoding for medical image segmentation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 11769-11779.

[38] Tang L, Deng Y, Yi X, et al. DRMF: Degradation-robust multi-modal image fusion via composable diffusion prior[C]//Proceedings of the 32nd ACM International Conference on Multimedia. 2024: 8546-8555.

[39] Yi X, Tang L, Zhang H, et al. Diff-IF: Multi-modality image fusion via diffusion model with fusion knowledge prior[J]. Information Fusion, 2024, 110: 102450.

[40] Xu R, Dong X M, Li W, et al. DBCTNet: Double branch convolution-transformer network for hyperspectral image classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.

[41] Qu L, Liu S, Wang M, et al. Transmef: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning[C]//Proceedings of the AAAI conference on artificial intelligence. 2022, 36(2): 2126-2134.

[42] Feng H, Wang L, Li Y, et al. LKASR: Large kernel attention for lightweight image super-resolution[J]. Knowledge-Based Systems, 2022, 252: 109376.

[43] James A P, Dasarathy B V. Medical image fusion: A survey of the state of the art[J]. Information fusion, 2014, 19: 4-19.