Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Unified Dual-Branch Framework for Tri-ModalMedical Image Fusion and Super-Resolution

Ye Liu
Linyi University
Yongli Chen
Linyi University
Ziyang Zhong
Linyi University
Ji Chen
Linyi University
Xing Wang
wangxing@lyu.edu.cn
Linyi University

Research Article

Keywords: Tri-modal medical image fusion, Medical multimedia analysis, DeformMamba, Dual-branch learning, Bidirectional cross-attention, Super-resolution

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9923301/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Unified Dual-Branch Framework for Tri-Modal Medical Image Fusion and Super-Resolution

Ye Liu[1], Yongli Chen[2], Ziyang Zhong[1], Ji Chen[1*], Xing Wang[1*]

[1]Information Science and Engineering, Linyi University, Linyi, China.
[2]Automation and Electrical Engineering, Linyi University, Linyi, China.

*Corresponding author(s). E-mail(s): chenji@lyu.edu.cn;
wangxing@lyu.edu.cn;
Contributing authors: liuye21@lyu.edu.cn; chenyongli_lyu@163.com;
zhongziyang_lyu@163.com;

Abstract

Multimodal medical imaging provides complementary anatomical and functional information for clinical diagnosis, yet low spatial resolution and separated fusion–super-resolution pipelines often weaken modality-specific details and reduce the quality of fused representations. To address this issue, this work proposes a unified dual-branch multimedia framework for tri-modal medical image fusion and super-resolution. The structure branch focuses on anatomical edges, contours, and local textures from structural modalities, while the function branch captures regional responses and semantic cues from functional modalities. A bidirectional cross-attention residual fusion module is further introduced to strengthen structural-functional interaction and improve complementary feature integration. The fused features are reconstructed under a joint loss that preserves both pixel fidelity and edge consistency. Experiments on the Harvard dataset show that the proposed method outperforms representative fusion and reconstruction baselines in quantitative evaluation and visual comparison. By jointly optimizing tri-modal fusion and super-resolution, the proposed framework provides high-quality fused medical representations for healthcare-oriented multimedia analysis, medical multimedia visualization, and subsequent image processing tasks.

Keywords: Tri-modal medical image fusion, Medical multimedia analysis, DeformMamba, Dual-branch learning, Bidirectional cross-attention, Super-resolution

1 Introduction

Multimodal medical imaging integrates complementary information from different imaging techniques and thereby provides comprehensive evidence for medical multimedia analysis tasks, ranging from lesion analysis to interactive clinical visualization [1]. For example, computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), and single-photon emission computed tomography (SPECT) reflect structural characteristics, functional status, and metabolic changes of human tissues from different perspectives, and recent medical image fusion studies have shown that multi-branch feature extraction and information exchange are effective for preserving modality-specific characteristics [2]. Among them, structural imaging modalities are more suitable for depicting anatomical structures, tissue boundaries, and local texture details, whereas PET and SPECT place greater emphasis on regional activity distribution, metabolic level, and functional response [3]. In modern computer-assisted medicine, high-quality multimodal representations are the cornerstone of Virtual Reality (VR)-based surgical planning and Augmented Reality (AR) navigation, and healthcare-oriented multimedia visualization [4, 5], where the visual fidelity of reconstructed images directly determines the precision of human-computer interaction and the reliability of clinical decisions. In clinical practice, physicians often need to jointly interpret structural and functional information to identify lesion locations, distinguish tissue boundaries, and observe abnormal metabolic responses. Retinal image-based deep learning has been explored for non-invasive diagnosis of diabetic kidney disease, and brain imaging-based deep learning systems have been developed for detecting silent brain infarction and predicting stroke risk [6, 7]. Therefore, a clear and high-resolution fused representation is valuable for lesion analysis, image-guided diagnosis, treatment planning, and subsequent medical image analysis. However, due to limitations in imaging devices and acquisition conditions, multimodal medical images often suffer from insufficient spatial resolution, blurred boundaries, and missing fine details in practice [8]. These issues reduce the readability of fused images and may affect subsequent analysis. Therefore, recovering clear structures and rich details while fully exploiting cross-modal complementary information remains an important problem for medical multimedia processing and healthcare-oriented visual analysis.

Existing studies related to this problem mainly follow two research lines, namely medical image fusion and medical image super-resolution reconstruction. Medical image fusion methods aim to integrate complementary information from different modalities to obtain more informative fused images [9]. For example, CDDFuse [10] decomposes multimodal features into low-frequency shared information and high-frequency modality-specific information through a correlation-driven dual-branch strategy, thereby improving the representation of common and modality-specific cues. EMMA [11] further introduces equivariant self-supervised learning to reduce the dependence on real fusion labels. SAFFusion [12] enhances structural details and functional hotspots from saliency-aware and frequency-domain perspectives, while AdaSFuse [13] explores adaptive cross-domain learning to improve the generalization ability of multimodal fusion. These methods have promoted multimodal feature representation and fusion quality. However, most existing fusion methods are still designed

for bimodal inputs or mainly validated in bimodal settings, and their ability to model tri-modal structural-functional complementarity remains limited. In addition, these methods usually focus on the generation quality of fused images themselves and do not directly address high-resolution fused image reconstruction from tri-modal inputs.

In parallel, medical image super-resolution methods focus on recovering high-frequency details and structural information from low-resolution medical images [14]. ICONet [15] uses reconstruction priors to iteratively guide medical image super-resolution, which helps improve local structure recovery. CuNeRF [16] constructs a continuous volumetric representation from a low-resolution medical volume itself and achieves zero-shot arbitrary-scale super-resolution. CycleINR [17] further leverages implicit neural representation with cycle-consistent constraints to reduce over-smoothing and inter-slice inconsistency. Although these methods improve the spatial resolution and visual quality of medical images, their main objective is usually to enhance a single target image or volume, making it difficult to fully exploit the complementary priors provided by multiple imaging modalities. As a result, fusion and super-resolution are still treated as relatively independent tasks in most studies, which results in irreversible information loss. Specifically, if image fusion is conducted prior to super-resolution reconstruction, modality-specific details may already be attenuated before reconstruction. Conversely, if super-resolution reconstruction is first applied to each individual modality followed by image fusion, reconstruction errors and inconsistent enhancement across different modalities are likely to be introduced.

Research that truly unifies medical image fusion and super-resolution within the same framework remains relatively limited. Xu et al. [18] proposed TFS-DIFF, which jointly models tri-modal medical image fusion and super-resolution within a conditional diffusion framework, providing an important reference for unified fusion reconstruction. Stephen et al. [19] discussed medical image fusion from the perspective of a diffusion-inspired multi-frame super-resolution framework, also emphasizing the connection between high-quality fused images and super-resolution reconstruction. Nevertheless, for tri-modal high-resolution fusion, simply placing fusion and reconstruction in the same framework is still insufficient. Structural and functional modalities differ clearly in information representation patterns and diagnostic emphases. Without targeted differentiated feature extraction and explicit structural-functional interaction, the obtained fused features may still be inadequate for effective high-resolution reconstruction.

To address these challenges, this paper proposes an end-to-end unified framework for tri-modal medical multimedia fusion and super-resolution, which integrates fusion and high-resolution reconstruction into a single optimization objective. By designing a dual-branch architecture, the model effectively bridges the gap between geometric anatomical details and semantic functional responses, providing a practical solution for healthcare-oriented multimedia analysis and medical image visualization.

The main contributions of this paper are summarized as follows:

1) An end-to-end joint optimization framework is proposed for tri-modal medical image fusion and super-resolution, enabling multimodal fusion and high-resolution reconstruction to be coordinated within a single model.

2) A structure branch and a function branch are designed to accommodate the representation differences between structural and functional modalities, and a bidirectional cross-attention residual fusion module, termed Bi-CARF, is introduced to enhance complementary modeling between structural details and functional responses.

3) Under unified experimental settings, the proposed method is compared with multiple representative methods, and both main and ablation experiments verify the effectiveness of the dual-branch representation, the bidirectional cross-modal fusion mechanism, and the related key configurations under different upscaling factors.

2 Related Work

2.1 Medical Image Fusion

Medical image fusion integrates complementary information from different imaging modalities. In traditional methods, Geng et al. [20] proposed an adaptive manifold filtering method that performs multimodal fusion through low-frequency and high-frequency decomposition together with an improved local contrast measure, achieving effective preservation of major structural information. Yin [21] proposed a tensor sparse representation based method for three-dimensional medical image fusion, which models volumetric data as a whole rather than processing it slice by slice, thereby enhancing the preservation of volumetric correlations. These methods have relatively clear modeling pathways and interpretability. However, their performance usually depends on predefined representation spaces and hand-crafted fusion rules, which limits their ability to model complex nonlinear cross-modal relationships. As fusion tasks extend from general multimodal fusion to more complex medical scenarios, this limitation becomes more evident.

Deep learning-based medical image fusion methods have been widely explored. Convolutional neural network (CNN)-based methods [22–24] first improved the representation ability of fusion models by replacing hand-crafted fusion rules with learnable feature extraction. For example, the supervised CNN fusion framework in [22] models multimodal inputs through convolutional encoding and batch-wise training, while VANet [23] introduces context feature extraction, an attention branch, and a multi-scale residual detail branch to enhance the diagnostic utility of fused images. In medical imaging scenarios, the interval-gradient and convolutional perceptual fusion method [24] further strengthens the hierarchical integration of structural, textural, and perceptual information. In addition to CNN-based models, generative adversarial network (GAN)-based methods introduce adversarial learning to improve visual realism and salient structure preservation. MedFusionGAN [25] adopts an unsupervised GAN framework for CT-MRI fusion, enhancing the joint representation of anatomical structures and soft-tissue contrast. More recently, diffusion-based methods have formulated multimodal fusion as a generative modeling problem. DDFM [26] uses a denoising diffusion model to improve multimodal image fusion, and the diffusion-driven medical fusion method of Qu et al. [27] further models the information distribution relationship among different modalities in the latent space. Transformer-based methods have also been introduced to strengthen local-global collaborative representation. SAT-Net [28] introduces a structure-aware Transformer-based attention fusion network

for low-quality retinal fundus image enhancement, showing that explicit structural modeling and attention-based fusion are effective for improving medical image quality. MDC-RHT [29] combines multidimensional dynamic convolution with a residual hybrid Transformer, whereas the edge-enhanced cross-scale Transformer framework [30] improves boundary preservation and multi-scale contextual modeling. Along the direction of efficient long-range dependency modeling, Mamba-based fusion methods have attracted increasing attention. FusionMamba [31] introduces dynamic feature enhancement and a cross-modal Mamba module for multimodal fusion, while ISFM [32] explores interactive spatial-frequency fusion based on the Mamba architecture. Overall, these methods have promoted medical image fusion from local convolutional feature learning to generative modeling, global dependency modeling, and efficient long-sequence representation. However, most of them still focus mainly on bimodal fusion, and their ability to support unified tri-modal fusion and high-resolution reconstruction remains limited.

For the tri-modal fusion task considered in this paper, related studies remain relatively limited. One relatively direct representative work is CTSR proposed by Jie et al. [33], which generates tri-modal fusion results through rolling guidance filtering, sparse representation, cartoon-texture decomposition, and an adaptive energy selection strategy. Existing tri-modal fusion studies still mainly focus on fused image generation itself, namely how to generate fused images with better visual quality, while paying limited attention to how the fused representation can further support subsequent high-resolution reconstruction.

2.2 Medical Image Super-Resolution

Medical image super-resolution recovers richer high-frequency details from the inputs, thereby improving spatial resolution, boundary discernibility, and clinical readability. Early studies mainly relied on interpolation, sparse representation, and prior constraints [34]. These methods usually compensate for missing information through explicit degradation modeling or external priors and have relatively clear optimization objectives. However, they are often limited in recovering complex textures and handling large upscaling factors.

With the development of deep learning, medical image super-resolution has gradually moved beyond simple pixel-wise regression toward stronger feature representation, perceptual reconstruction, and efficient dependency modeling. CNN-based methods [35–37] first improved the recovery of local structures and fine details through learnable feature extraction, residual connections, and dense feature reuse. Dai et al. [35] proposed a second-order attention network, which captures second-order statistical information among features to improve reconstruction quality. Zhang et al. [36] then introduced a deep residual dense network to further enhance feature reuse, effectively improving the clarity of anatomical structures. Med-SRNet [37] further strengthens detail recovery through high-resolution representation learning. However, pure CNN-based methods are still limited in capturing global dependencies and generating realistic textures, which has motivated researchers to explore other methods. GAN-based methods introduce adversarial optimization to enhance visual realism and texture perception. Ahmad et al. [38] proposed a GAN-based framework to

improve texture realism and visual quality, while Zhang et al. [39] incorporated long short-term memory and attention into a GAN-based framework for three-dimensional medical image super-resolution, thereby exploiting sequential correlations across volumetric slices. Transformer-based methods further strengthen global dependency modeling and efficient attention computation. SG-SRNet [40] computes attention only in information-dense regions through sparse masking and clustered attention, reducing redundant computation in multimodal medical image super-resolution. Qin et al. [41] further proposed a CNN–Transformer gated fusion network, which combines the global modeling ability of Transformers with the local detail restoration ability of dynamic convolution. Diffusion-based methods have also been introduced to improve medical image reconstruction quality and processing efficiency. Namatevs et al. [42] reduced denoising steps by comparing the intermediate distributions of low- and high-resolution images, while Sadikov et al. [43] adopted Swin UNETR for efficient denoising of diffusion MRI. Both methods significantly improve processing speed while maintaining reconstruction quality, providing a feasible path toward clinical deployment. In recent years, Mamba architectures have also been introduced into medical image super-resolution. GLMamba [44] adopts a global Mamba branch and a local Mamba branch to model long-range dependencies and local details, respectively, and integrates complementary information through a multimodal feature fusion block. Overall, medical image super-resolution has evolved from local convolutional reconstruction and adversarial perceptual enhancement toward global dependency modeling, diffusion-based generative modeling, and efficient long-sequence representation. Nevertheless, most existing studies still focus on single-modality reconstruction or reference-guided multimodal super-resolution, and their final objective is usually to enhance the resolution of a specific target image. Few studies treat fusion and super-resolution as a unified objective for generating a high-resolution fused representation from multiple complementary modalities.

3 Method

3.1 Overall Framework

From the perspective of medical imaging characteristics, different modalities emphasize different types of information. Structural images represented by modalities such as T1-weighted MRI (T1) and T2-weighted MRI (T2) can more clearly depict tissue boundaries, morphological contours, and local texture details, which are closely related to structural preservation in fusion results and subsequent high-resolution reconstruction. In contrast, functional images such as SPECT are more strongly associated with regional activity distribution, global intensity variation, and functional response in target regions. Because these two types of features differ in both representation patterns and points of emphasis, simple fusion is often insufficient to fully exploit their complementary advantages. To address this issue, this paper designs a dual-branch network architecture, as shown in Fig. 1, to better explore the complementary properties of different modalities. Specifically, a structure branch is designed to encode inputs with strong structural information in a targeted manner, so as to extract stable priors of edges, contours, and local textures. At the same time, a function branch

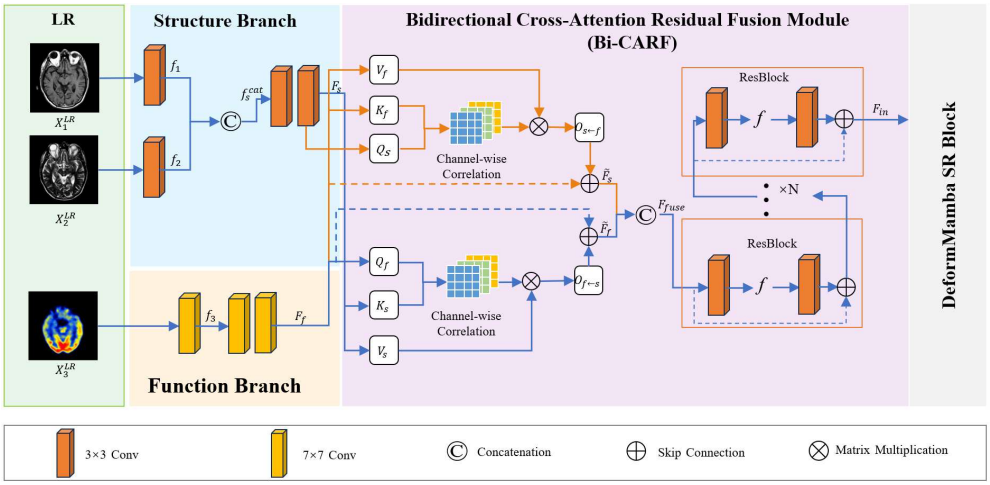


Fig. 1 Overall architecture of the proposed unified framework for tri-modal medical image fusion and super-resolution. The network adopts an end-to-end pipeline consisting of three core modules: a dual-branch differentiated feature extractor (with a structure branch for anatomical detail extraction and a function branch for semantic response capture), a bidirectional cross-attention residual fusion module, and a high-resolution reconstruction head optimized under joint loss constraints. LR and SR denote low-resolution input images and super-resolution outputs, respectively.

is introduced to extract functional features with stronger contextual perception ability, thereby providing the necessary semantic complement for subsequent cross-modal fusion. To further enhance the interaction and fusion between these two types of features, a Bi-CARF module is built on top of the dual-branch architecture, where bidirectional cross-modal interaction is used to achieve effective complementarity between structural details and functional semantics.

3.2 Structure Branch

In practical applications, structural modalities may involve multiple types, such as T1, T2, and contrast-enhanced T1-weighted images, and the proposed method can be naturally extended to an arbitrary number of structural inputs. For ease of presentation, we take two representative structural modalities, X_1^{LR} and X_2^{LR}, as examples. Since the informative content of structural modalities is mainly concentrated in local edges, contour boundaries, and fine-grained textures, the structure branch first employs two independent 3×3 convolutions to extract the initial structural features $f_1 \in \mathbb{R}^{C \times H \times W}$ and $f_2 \in \mathbb{R}^{C \times H \times W}$. This step enables a preliminary extraction of local structural responses while preserving the original imaging characteristics, where C, H, and W denote the number of channels, height, and width of the feature maps, respectively. The resulting features are then concatenated along the channel dimension to obtain the structural feature $f_s^{cat} \in \mathbb{R}^{2C \times H \times W}$. This strategy, which first encodes each modality independently and then concatenates the features, avoids premature mixing of different structural modalities at the shallow stage and preserves more sufficient modality-specific information for subsequent feature combination. Based on this

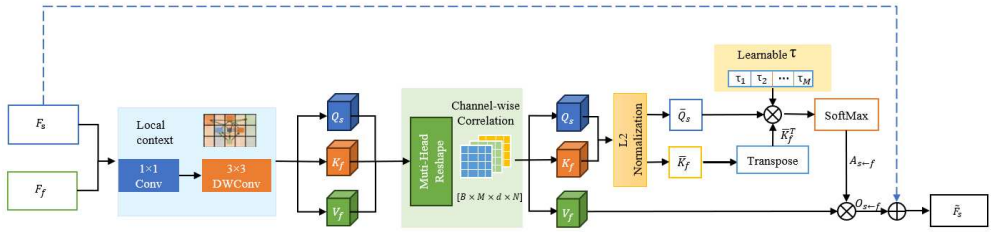


Fig. 2 One directional path of the proposed Bi-CARF. The structure-guided aggregation of functional features is shown as an example.

representation, f_s^{cat} is fed into a structure encoder composed of two 3×3 convolutions and LeakyReLU activation functions. The first convolution mainly models local correlations between the concatenated bimodal features and performs channel compression. The second convolution further refines the structural representation under a unified feature dimension, allowing boundary and texture information to be more fully integrated. The final output of the structure branch is $F_s \in \mathbb{R}^{C\times H\times W}$.

3.3 function branch

Similar to the structure branch, the functional modality may also involve multiple input types, such as PET and SPECT, and the proposed method can likewise be extended to multiple functional inputs. For ease of presentation, we take a single representative functional modality, X_3^{LR}, as an example. Since the effective information in the functional modality is mainly reflected in regional responses over relatively large spatial ranges, the function branch first applies a 7×7 convolution for initial mapping to obtain an initial functional feature $f_3 \in \mathbb{R}^{C\times H\times W}$. On this basis, f_3 is then fed into a functional encoder for further feature extraction. The encoder consists of two 7×7 convolution layers and LeakyReLU activation functions, which continuously aggregate functional responses within large spatial neighborhoods and model contextual correlations across different regions, thereby improving the stability and completeness of functional information representation. The final output is the function branch feature $F_f \in \mathbb{R}^{C\times H\times W}$.

3.4 Bidirectional Cross-Attention Residual Fusion Module (Bi-CARF)

After obtaining the structure branch feature $F_s \in \mathbb{R}^{B\times C\times H\times W}$ and the function branch feature $F_f \in \mathbb{R}^{B\times C\times H\times W}$, the key objective of the fusion stage is to establish a stable correspondence between these two types of features. Since they focus on different image attributes, direct addition or concatenation often implicitly assumes that cross-modal features are already well aligned in spatial location, whereas this assumption is difficult to satisfy in medical images. To address this issue, this paper designs a bidirectional cross-attention residual fusion module, namely Bi-CARF, which achieves interactive fusion between structural and functional features through a cross-modal attention mechanism. Furthermore, given the strong complementarity between the

two feature types in information representation, one-way guidance is insufficient to fully model their mutual dependencies. Therefore, a bidirectional interaction strategy is adopted. Since the computations in the two directions are essentially the same, only one interaction path is presented to avoid redundancy. As illustrated in Fig. 2, the structural feature is used as the query to select responses from the functional feature that are relevant to detail restoration. Specifically,

$$\begin{aligned} Q_s &= D_q^s\big(W_q^s(F_s)\big) \\ K_f &= D_k^f\big(W_k^f(F_f)\big) \\ V_f &= D_v^f\big(W_v^f(F_f)\big) \end{aligned} \tag{1}$$

where $W(\cdot)$ denotes a 1×1 convolution and $D(\cdot)$ denotes a 3×3 depthwise convolution (DWConv).This formulation generates the queries, keys, and values required for cross-attention through linear projections. Introducing DWConv after the linear projection supplements local neighborhood information before attention computation, thereby improving the stability of cross-modal matching.

The features are then rearranged into multi-head form as $Q_s, K_f, V_f \in \mathbb{R}^{B\times M\times d\times N}$, where M is the number of attention heads, $d = C/M$, and $N = H \times W$. Since the reshaped Q and K have dimensions $B \times M \times d \times N$, $\bar{Q}_s\bar{K}_f^\top$ produces a channel correlation matrix rather than a conventional spatial correlation matrix. This design allows structural channels to extract information from better-matched functional channels, while avoiding the additional computational cost introduced by spatial attention maps for high-resolution features. On this basis, Q_s and K_f are normalized along the last dimension to reduce the interference caused by differences in intensity distribution across modalities. The attention weight matrix for the structure-guided path is defined as:

$$A_{s\leftarrow f} = \mathrm{Softmax}\big(\tau\,\bar{Q}_s\bar{K}_f^\top\big) \tag{2}$$

where $A_{s\leftarrow f}$ denotes the attention weight matrix for the structure branch to aggregate information from the function branch, τ is a learnable parameter for different attention heads. Considering that different attention heads may focus on different feature relationships during cross-modal interaction, a fixed scaling factor cannot flexibly satisfy the attention distribution required by each head. Therefore, τ is set as a learnable parameter, allowing each head to adaptively adjust the aggregation strength. The resulting cross-modal output is expressed as:

$$O_{s\leftarrow f} = A_{s\leftarrow f}V_f \tag{3}$$

where $O_{s\leftarrow f}$ represents the output feature of this path.

The module then updates the original structural feature in a residual manner to obtain the enhanced representation $\tilde{F}_s \in \mathbb{R}^{B\times C\times H\times W}$. The reverse path is defined in the same way, yielding the enhanced functional feature $\tilde{F}_f$. Finally, the enhanced features from both directions are concatenated along the channel dimension to obtain the fused feature $F_{fuse} \in \mathbb{R}^{B\times 2C\times H\times W}$. Since bidirectional interaction makes the feature distribution more complex, a lightweight residual refinement module $R(\cdot)$ is further introduced after F_{fuse} to locally reorganize the fused result and produce the

feature F_{in}. Finally, F_{in} is fed into the reconstruction network to complete super-resolution.

3.5 Loss Function

For multimodal medical image super-resolution, relying only on pixel-level error can make the overall intensity of the reconstruction close to that of the real image, but it remains insufficient for recovering edge contours and local high-frequency structures. To address this issue, this paper adopts a joint loss function that combines a pixel reconstruction constraint with an edge consistency constraint, so as to improve both the overall fidelity and the detail restoration capability of the reconstructed results.

Specifically, a pixel reconstruction loss is used to directly constrain the difference between the network output and the ground-truth high-resolution image. Here, the L_1 loss is adopted and defined as:

$$\mathcal{L}_{pix} = \|SR - HR\|_1 \tag{4}$$

where SR denotes the network output and HR denotes the ground-truth high-resolution image.

In addition, to further enhance the model's sensitivity to tissue boundaries and local high-frequency structures, an edge consistency loss is introduced. Specifically, after converting the predicted image and the ground-truth image to grayscale, three fixed 3×3 edge extraction kernels are used to compute their local differential responses. Let the response of the m-th edge operator be denoted by $E^m(SR)$ and $E^m(HR)$, respectively. The edge consistency loss is then defined as:

$$\mathcal{L}_{edge} = \frac{1}{3}\sum_{m=1}^{3} \|E^m(SR) - E^m(HR)\|_1 \tag{5}$$

This loss term constrains the consistency between the prediction and the ground truth in terms of local gradients and boundary responses, thereby encouraging the network to better preserve tissue boundaries, structural contours, and fine-grained texture information while restoring the overall intensity distribution. By combining the two constraints above, the total loss function is written as:

$$\mathcal{L} = \mathcal{L}_{pix} + \beta\mathcal{L}_{edge} \tag{6}$$

where β is a weighting coefficient used to balance the contributions of the pixel reconstruction loss and the edge consistency loss. Different weight settings are compared in preliminary experiments, and $\beta = 0.1$ is finally selected after jointly considering pixel fidelity and structural detail restoration. This setting maintains the overall intensity accuracy of the reconstruction while further strengthening the constraint on high-frequency edge information, thereby yielding a better overall optimization effect.

4 Experiment

4.1 Experimental Details

The proposed method is implemented in the PyTorch framework and is trained and tested on a workstation equipped with an NVIDIA GeForce RTX4090 GPU. During training, the Adam optimizer is used with a learning rate of 1×10^{-4}, a batch size of 4, and a total of 200 epochs. The same training strategy is adopted in all experiments to minimize the influence of irrelevant factors on the results. Following the dataset setting in the existing TFS-DIFF [18] work, all competing methods and the proposed method are evaluated under the same data split and the same training and test data to ensure the fairness and comparability of the experimental results. Since the task considered in this study simultaneously emphasizes overall intensity recovery, edge preservation, and structural detail fidelity, peak signal-to-noise ratio (PSNR) [45] and structural similarity index measure (SSIM) [45] are adopted as evaluation metrics. PSNR reflects pixel-level reconstruction fidelity, with a higher value indicating lower distortion from the reference image, while SSIM measures perceptual structural similarity in terms of luminance, contrast, and structure, with a value closer to 1 indicating better structural preservation.

To validate the effectiveness of the proposed method, several representative medical image fusion and reconstruction methods from recent years are selected for comparison, including CTSR [33], CDDFuse [10], DDFM [26], EMMA [11], TFS-DIFF [18], FusionMamba [31], ISFM [32], AdaSFuse [13], EFRUFN [46] and Mask-DiFuser [47]. For all competing methods, reproduction is conducted based on their publicly available code and recommended parameter settings. Since most of these methods are originally designed for bimodal fusion, whereas this study focuses on joint tri-modal fusion and super-resolution, the input stage is extended to three modalities in the experiments without changing their core fusion mechanisms, and a unified DeformMamba [48] reconstruction module is combined to perform super-resolution. The purpose of this design is to constrain the differences among methods, as much as possible, to the fusion mechanism itself and to reduce the interference caused by differences in reconstruction modules, thereby improving the reasonableness of the comparison.

4.2 Quantitative Comparison Results

Table 1 presents the quantitative results of all compared methods on the Harvard test set. PSNR and SSIM are reported at the $\times 2$, $\times 4$, and $\times 8$ scale factors to comprehensively evaluate the reconstruction accuracy and structural preservation capability of different methods.

At the $\times 2$ scale, Mask-DiFuser, ISFM, TFS-DIFF, AdaSFuse, CDDFuse, and EMMA show relatively competitive performance among the compared methods, whereas DDFM obtains clearly lower results. The proposed method achieves a PSNR of 28.94 dB and an SSIM of 0.93, obtaining the highest PSNR and the tied-best SSIM. Compared with Mask-DiFuser, which is the strongest competing method at this scale, the proposed method improves PSNR by 0.88 dB while maintaining the same SSIM

Table 1 Quantitative comparison on the Harvard dataset at ×2, ×4, and ×8 upscaling factors. PSNR and SSIM are reported for all methods, where higher values indicate better performance. The best result in each column is shown in bold. All compared methods use the same DeformMamba reconstruction module for super-resolution.

	×2		×4		×8	
Method	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
CTSR [33] (TIM'23)	25.73	0.86	23.05	0.78	20.84	0.72
CDDFuse [10] (CVPR'23)	26.55	0.89	22.38	0.77	20.15	0.70
DDFM [26] (ICCV'23)	20.55	0.74	19.31	0.69	18.49	0.66
EMMA [11] (CVPR'24)	26.49	0.88	23.03	0.79	20.18	0.70
TFS-DIFF [18] (MICCAI'24)	27.16	0.89	24.02	0.81	21.60	0.74
FusionMamba [31] (VI'25)	25.33	0.86	22.99	0.79	20.92	0.72
ISFM [32] (TIP'25)	27.91	0.91	24.13	0.85	20.07	0.70
AdaSFuse [13] (TMM'25)	26.74	0.88	23.50	0.79	21.20	0.73
EFRUFN [46] (TPAMI'26)	26.71	0.90	22.56	0.80	19.42	0.66
Mask-DiFuser [47] (TPAMI'26)	28.06	**0.93**	24.36	**0.88**	21.39	**0.79**
Ours	**28.94**	**0.93**	**24.95**	0.84	**21.93**	**0.77**

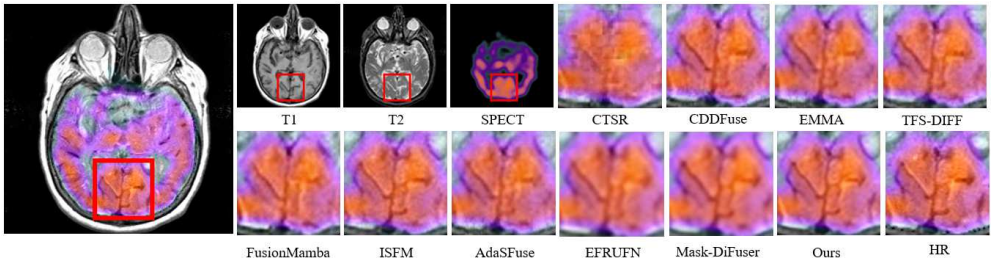


Fig. 3 Visual comparison of tri-modal fusion and ×2 super-resolution on T1, T2, SPECT images.

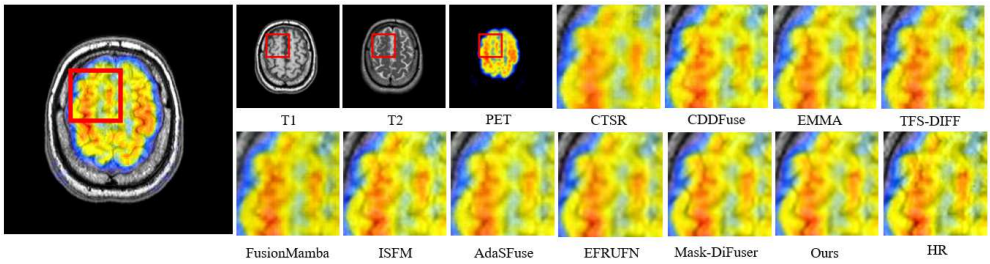


Fig. 4 Visual comparison of tri-modal fusion and ×2 super-resolution on T1, T2, PET images.

value. Compared with ISFM and TFS-DIFF, the proposed method further improves PSNR by 1.03 dB and 1.78 dB, respectively.

When the scale factor increases to ×4, the proposed method still achieves the highest PSNR of 24.95 dB, outperforming Mask-DiFuser, ISFM, and TFS-DIFF by 0.59 dB, 0.82 dB, and 0.93 dB, respectively. In terms of SSIM, Mask-DiFuser obtains the best value of 0.88, while the proposed method achieves 0.84. Although the SSIM of the proposed method is slightly lower than that of Mask-DiFuser, its clear advantage in PSNR indicates stronger intensity reconstruction fidelity and lower pixel-level distortion under this scale.

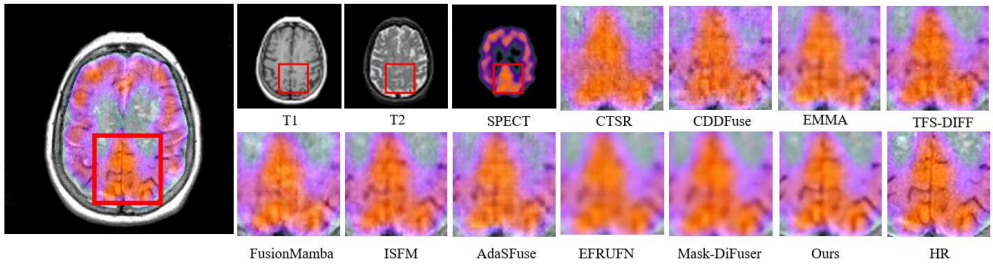


Fig. 5 Visual comparison of tri-modal fusion and ×4 super-resolution on T1, T2, and SPECT images.

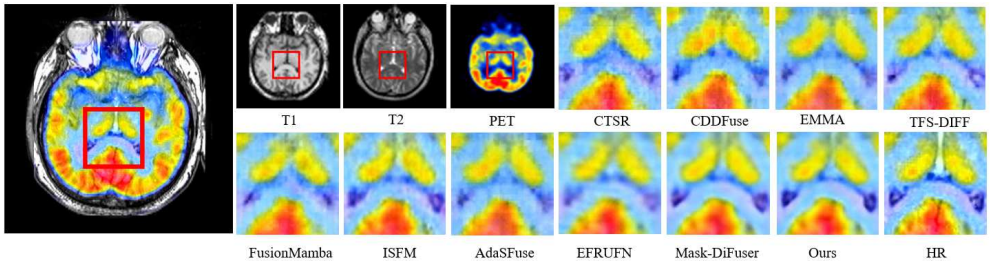


Fig. 6 Visual comparison of tri-modal fusion and ×4 super-resolution on T1, T2, and PET images.

At the ×8 scale, the performance degradation of most methods becomes more evident due to the severe loss of high-frequency information. The proposed method achieves the highest PSNR of 21.93 dB, surpassing TFS-DIFF, Mask-DiFuser, and AdaSFuse by 0.33 dB, 0.54 dB, and 0.73 dB, respectively. Mask-DiFuser obtains the highest SSIM of 0.79, while the proposed method achieves the second-best SSIM of 0.77. This result suggests that Mask-DiFuser has an advantage in structural similarity preservation, whereas the proposed method provides better overall reconstruction fidelity. Therefore, the proposed method achieves a more favorable balance between pixel-level accuracy and structural preservation, especially under challenging high-magnification reconstruction conditions.

4.3 Visual Comparison Analysis

In addition to the quantitative comparison, this paper further selects several representative competing methods for visual analysis of reconstruction quality. The results at the ×2, ×4, and ×8 upscaling factors are shown in Fig. 3–8, respectively. The left side of each figure presents the overall fused image, with the local comparison region marked by a red box. The right side shows magnified views of this region for T1, T2, SPECT (PET), and different methods, where HR denotes the high-quality reference image. As shown in Fig. 3 and 4, at the ×2 upscaling factor, most compared methods can recover the main anatomical structures, but noticeable differences remain in the magnified local regions. CTSR and CDDFuse preserve the coarse structural layout, but some fine textures and local boundaries are still over-smoothed. EMMA, TFS-DIFF,

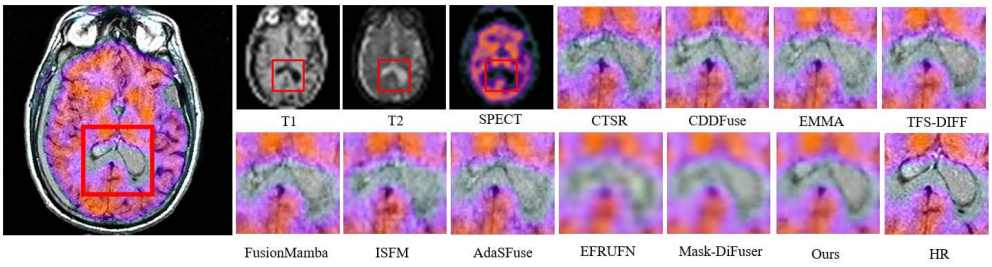


Fig. 7 Visual comparison of tri-modal fusion and ×8 super-resolution on T1, T2, SPECT images.

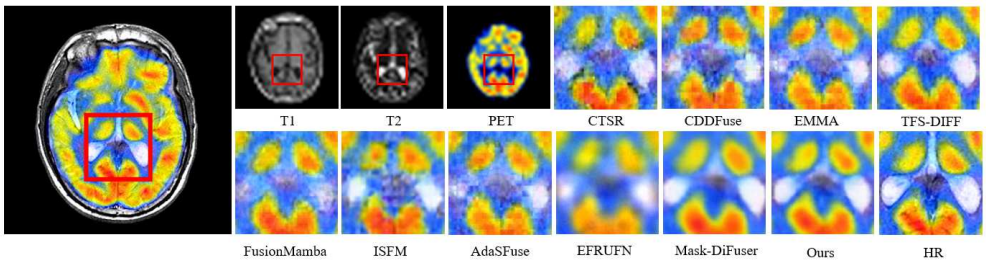


Fig. 8 Visual comparison of tri-modal fusion and ×8 super-resolution on T1, T2, PET images.

and FusionMamba recover approximate texture patterns, while texture adhesion and blurred transitions can still be observed in several fine-detail regions. ISFM and Mask-DiFuser show relatively competitive structural preservation, which is consistent with their quantitative SSIM performance. However, their local details are still not as clear as those of the HR reference. In contrast, the proposed method reconstructs finer branch-like textures and maintains clearer boundary transitions in both SPECT and PET related results, producing visual details that are closer to the HR images.

As shown in Fig. 5 and 6, at the ×4 upscaling factor, different methods exhibit clear differences in their ability to preserve local structures and functional information. The red-boxed region in Fig. 5 mainly focuses on the central orange functional structure and its internal vertical textures. Methods such as CTSR, EMMA, EFRUFN, and Mask-DiFuser show insufficient detail recovery, with relatively blurred structural boundaries and internal textures. Although CDDFuse, FusionMamba, ISFM, and AdaSFuse preserve the approximate morphology, their texture continuity and structural stability remain limited. The red-boxed region in Fig. 6 mainly contains a central dark-blue low-response structure, bilateral yellow functional distributions, and a lower red-orange high-response area. Although most comparison methods can recover the overall color distribution, they are still prone to over-smoothing, blurred boundaries, or attenuation of the dark-blue structure, resulting in insufficient local structural hierarchy. In contrast, By comparison, the proposed method better preserves the intensity distribution of functional regions in both cases, while recovering clearer boundaries, local textures, and structural hierarchies.

Table 2 Ablation study of different module combinations. ✔ and ✗ denote whether the corresponding structure branch, function branch, or fusion module is used. Concat denotes direct channel concatenation, and Bi-CARF denotes the proposed bidirectional cross-attention residual fusion module.

Structure Branch	Function Branch	Fusion module	×2		×4		×8	
			PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
✔	✗	✗	21.51	0.82	20.60	0.75	19.11	0.68
✗	✔	✗	17.27	0.66	17.43	0.66	17.66	0.65
✔	✔	Concat	28.14	0.91	24.60	0.84	21.77	0.75
✔	✔	Bi-CARF	**28.94**	**0.93**	**24.95**	**0.84**	**21.93**	**0.77**

As shown in Fig. 7 and 8, at the ×8 upscaling factor, the information loss caused by large-magnitude enlargement becomes more pronounced, and the visual differences among methods are mainly reflected in their ability to preserve key discriminative structures. In the red-boxed regions, the main differences lie in the continuity of the central low-response area, its separation from the surrounding high-response regions, and the boundary integrity between adjacent anatomical structures and the background. It can be seen that CTSR, CDDFuse, EMMA, and AdaSFuse are limited in recovering the central low-response structure. Typical artifacts include blunted contours, spatial expansion of local regions, and unclear boundaries between adjacent response areas. EFRUFN produces noticeably over-smoothed results, where several local structures become weakened and difficult to distinguish. TFS-DIFF and FusionMamba better preserve structural relationships in the vertical and horizontal directions, but the continuity of the central low-response region remains insufficient, and the boundary transitions are still overly broad. ISFM presents more local variations, but its internal textures appear somewhat disordered, and the separation between regions is not sufficiently stable. Mask-DiFuser shows relatively good structural preservation, which is consistent with its SSIM performance, but its local pixel-level details are less faithful to the HR reference in some regions. In comparison, the proposed method maintains a more discernible central structure and relatively clearer separation between adjacent regions, while preserving comparatively complete local contours in both comparisons. Nevertheless, some smoothing can still be observed compared with the HR images, indicating that further improvement is still needed for fine-detail recovery under severe upsampling conditions.

4.4 Ablation Experiment

To verify the effectiveness of the proposed dual-branch structure and cross-modal fusion strategy, this paper conducts the main ablation experiments to compare the reconstruction performance of different module combinations at three upscaling factors. Specifically, four model variants are considered, including a structure-only branch, a function-only branch, dual branches with direct concatenation (Concat), and the full model that combines dual branches with Bi-CARF. The detailed configurations of these variants are listed in Table 2. In the ablation experiments, all training parameters and training strategies are kept identical across settings except for the module combination scheme, in order to ensure the fairness and comparability of the results.

Table 3 Effect of the loss-weight coefficient β.

	×2		×4		×8	
β	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
0.1	**28.94**	**0.93**	**24.95**	**0.84**	**21.93**	**0.77**
0.2	28.56	0.91	24.85	0.84	21.82	0.75
0.5	28.60	0.92	24.92	0.84	21.86	0.76
0.8	28.57	0.92	24.90	0.84	21.86	0.75
1.0	28.62	0.92	24.86	0.84	21.69	0.75

To verify the necessity of the dual-branch design, the dual-branch direct concatenation is compared with the two single-branch settings. As shown in Table 2, compared with using only the structure branch, Concat improves PSNR by 6.63 dB, 4.00 dB, and 2.66 dB at the ×2, ×4, and ×8 upscaling factors, respectively, while SSIM increases by 0.09, 0.09, and 0.07. Compared with using only the function branch, Concat improves PSNR by approximately 10.87 dB, 7.17 dB, and 4.11 dB at the ×2, ×4, and ×8 upscaling factors, respectively, while SSIM increases by approximately 0.25, 0.18, and 0.10. These results indicate that a single branch is insufficient to simultaneously account for structural detail recovery and functional semantic representation, whereas the dual-branch design effectively integrates the complementary advantages of different modalities, thereby substantially improving image reconstruction performance and confirming the necessity of the dual-branch architecture.

Based on the dual-branch architecture, this paper further compares simple concatenation with the variant equipped with Bi-CARF. Compared with Concat, Bi-CARF improves PSNR by 0.8 dB, 0.35 dB, and 0.16 dB at the three upscaling factors, respectively. For SSIM, it yields an improvement of 0.02 at both the ×2 and ×8 upscaling factors, while remaining unchanged at the ×4 factor. These results suggest that, once dual-branch features are available, simple concatenation alone is insufficient to establish effective cross-modal correspondence. In contrast, Bi-CARF further explores the correlation between structural and functional information through bidirectional interaction, thereby enabling more adequate feature fusion.

4.5 Additional Analysis

To further examine the effects of several key design details on model performance, this section provides a detailed analysis from three aspects, namely hyperparameter setting, the internal structure of the fusion module, and convolution kernel size selection. By comparing the experimental results under different configurations, the rationale of these design choices can be more clearly demonstrated. The corresponding results are summarized in Tables 3–5. To evaluate the effect of the edge-consistency loss weight coefficient β on model performance, β is set to 0.1, 0.2, 0.5, 0.8, and 1.0, and the results are compared at the three upscaling factors of ×2, ×4, and ×8, as shown in Table 3. Overall, model performance exhibits only slight fluctuations as β varies, indicating that the proposed method remains relatively robust to the choice of this hyperparameter within a certain range. More specifically, when $\beta = 0.1$, the model achieves the best PSNR at all three upscaling factors and obtains the best or tied-best SSIM values. However, when β becomes excessively large, performance decreases slightly. Considering the results across all three upscaling settings, β is ultimately set to 0.1 in this paper.

Table 4 Internal ablation study of Bi-CARF. Different interaction directions, together with the use or removal of DWConv and normalization, are compared. S and F denote structural and functional features, respectively. S ← F denotes the structure-guided aggregation path, F ← S denotes the reverse path.

			×2		×4		×8	
Direction	DWConv	Normalization	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Single(F ← S)	✔	✔	28.02	0.90	24.10	0.81	21.75	0.74
Single(S ← F)	✔	✔	28.21	0.91	24.73	0.84	21.78	0.75
Bidirectional	✗	✔	27.89	0.90	24.55	0.82	21.30	0.73
Bidirectional	✔	✗	28.20	0.91	24.73	0.84	21.78	0.75
Bidirectional	✔	✔	**28.94**	**0.93**	**24.95**	**0.84**	**21.93**	**0.77**

Table 5 Effect of convolution kernel size in the structure and function branches.

		×2		×4		×8	
Structure kernel	Function kernel	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
7×7	7×7	27.86	0.90	24.53	0.81	21.60	0.74
3×3	3×3	28.24	0.92	24.80	0.83	21.81	0.75
3×3	7×7	**28.94**	**0.93**	**24.95**	**0.84**	**21.93**	**0.77**

To further analyze the effects of the key components in Bi-CARF on fusion and reconstruction performance, internal ablation experiments are conducted from three aspects, namely interaction direction, the local modeling unit, and feature normalization, and the results are presented in Table 4. Compared with the two unidirectional interaction variants, the complete bidirectional interaction achieves the highest PSNR at all upscaling factors and obtains comparable or better SSIM values. This indicates that bidirectional interaction is more effective in establishing the mutual relationship between the two types of features. After removing DWConv, the PSNR drops by 1.05 dB, 0.40 dB, and 0.63 dB at the three upscaling factors, respectively. In addition, removing the normalization operation also leads to performance degradation to varying degrees. Overall, the performance gain of Bi-CARF does not arise from any single operation, but from the joint effect of bidirectional interaction, local modeling, and normalization constraints.

To verify the rationality of the convolution kernel settings in the structure branch and the function branch, three kernel combinations are considered: using 7×7 convolutions in both branches, using 3×3 convolutions in both branches, and using a 3×3 convolution in the structure branch together with a 7×7 convolution in the function branch. As shown in Table 5, the model achieves the best performance when different kernel sizes are adopted in the two branches. This indicates that the structural and functional modalities emphasize different aspects of information representation, and that using different kernel sizes is more beneficial for extracting effective features from each modality.

5 Discussion

Although the unified framework proposed in this paper achieves favorable overall performance in tri-modal medical image fusion and super-resolution reconstruction, several limitations remain to be further optimized and improved for real-world clinical applications.

The current framework assumes that the input multimodal images are spatially well registered. However, in clinical practice, multimodal medical images are often acquired using different devices, scanning protocols, and time points. Patient motion and variations in imaging parameters may further introduce local spatial misalignments. Such registration errors may disrupt the correspondence between structural information and functional responses, interfere with cross-modal attention-based feature learning, and ultimately degrade the quality of image fusion and super-resolution reconstruction. Further, the proposed method assumes access to complete tri-modal inputs during inference. In real clinical environments, however, complete tri-modal data are not always available due to constraints such as scanning cost, radiation dose, examination duration, patient contraindications, and incomplete clinical records. Missing modalities or severely degraded image quality are common in practical scenarios, which reveals the limited adaptability and stability of the current dual-branch structure under incomplete input conditions. In addition, the model mainly relies on supervised training with pseudo-ground-truth references. Although this training setting enables consistent experimental evaluation and comparable quantitative results, obtaining large amounts of strictly paired and high-quality ground truth for tri-modal high-resolution fused images remains challenging in medical imaging. This may limit the practical deployment potential of the model to some extent. Moreover, the model has only been validated on a single dataset, and its generalization ability across different imaging devices, scanning parameters, anatomical regions, and clinical centers still requires further systematic verification.

To address these limitations, future work will conduct systematic research focusing on the clinical applicability, robustness, and generalizability of the model. For registration errors commonly observed in clinical settings, future studies will prioritize improving model stability under non-ideal registration conditions and explore fusion and reconstruction strategies that can adapt to complex spatial misalignment. For incomplete modality data and degraded image quality in real clinical environments, future research will focus on missing-modality robust learning to improve the model's adaptability and robustness to diverse and incomplete inputs. To reduce the reliance on high-quality paired ground truth, future work will explore learning paradigms such as unsupervised and weakly supervised learning, thereby reducing the dependence on precise annotations and better adapting to the scarcity and pairing difficulty of medical imaging data. In addition, to further promote the clinical translation of the model, future research will emphasize improving cross-scenario generalization through validation and optimization on large-scale, multi-center, and diverse datasets. The clinical potential and practical value of the proposed fusion and reconstruction framework will also be comprehensively evaluated through clinical expert assessment and downstream diagnostic tasks.

6 Conclusion

This paper presents a unified framework for tri-modal medical image fusion and super-resolution reconstruction. The proposed framework performs differentiated modeling of multimodal information through a structure branch and a function branch, and

explicitly establishes the relationship between structural information and functional semantics via a bidirectional cross-attention residual fusion module. In this way, end-to-end modeling from tri-modal inputs to high-resolution fused outputs is achieved. Extensive experiments on multiple tri-modal medical image combinations demonstrate that the joint optimization of image fusion and super-resolution reconstruction effectively improves both fusion quality and reconstruction performance, and that the proposed framework achieves superior overall performance. This paper also discusses the potential limitations of the proposed method and future work in real-world clinical applications.

Acknowledgements This work was supported in part by the Natural Science Foundation of Shandong Province under Grant ZR2024QF201, the National Natural Science Foundation of China under Grant 62341603, and the Introduction and Cultivation Program for Young Innovative Talents of Universities in Shandong Province under Grant 2021QCYY003.

References

[1] Zubair, M., Hussain, M., Albashrawi, M.A., *et al.*: A comprehensive review of techniques, algorithms, advancements, challenges, and clinical applications of multi-modal medical image fusion for improved diagnosis. Computer Methods and Programs in Biomedicine **272**, 109014 (2025). https://doi.org/10.1016/j.cmpb.2025.109014

[2] Ravirathinam, P., Goel, D., Ranjani, J.J.: C-lienet: A multi-context low-light image enhancement network. IEEE Access **9**, 31053–31064 (2021). https://doi.org/10.1109/ACCESS.2021.3059498

[3] Wu, K., Guan, S., Pu, J., Cheng, N., Lin, X., Jiang, N.: Cipher: cross-modal integrated and pathology-aware embedded representation diffusion model with temporal consistency for brain mri-to-pet translation. The Visual Computer **42**, 248 (2026). https://doi.org/10.1007/s00371-026-04438-6

[4] Tang, H., Yu, L., Shao, Y.: Marcfusion: adaptive residual cross-domain fusion network for medical image fusion. Multimedia Systems **31**, 31053–31064 (2025). https://doi.org/10.1007/s00530-025-01686-9

[5] Ha, T.L., Verde, J., Bert, J., et al.: Toward fluoroscopy guided robotic needle insertion for radio frequency ablation. Computer Animation and Virtual Worlds **36** (2025). https://doi.org/10.1002/cav.70025

[6] Meng, Z., Guan, Z., Yu, S., *et al.*: Non-invasive biopsy diagnosis of diabetic kidney disease via deep learning applied to retinal images: a population-based study. The Lancet. Digital health **7**, 100868 (2025). https://doi.org/10.1016/j.landig.2025.02.008

[7] Jiang, N., Ji, H., Guan, Z., *et al.*: A deep learning system for detecting silent brain infarction and predicting stroke risk. Nature Biomedical Engineering **9**, 1907–1919 (2025). https://doi.org/10.1038/s41551-025-01413-9

[8] Umirzakova, S., Ahmad, S., Khan, L.U., *et al.*: Medical image super-resolution for smart healthcare applications: A comprehensive survey. Information Fusion **103**, 102075 (2023). https://doi.org/10.1016/j.inffus.2023.102075

[9] Rajalingam, B., Al-turjman, F.M., Santhoshkumar, R., *et al.*: Intelligent multi-modal medical image fusion with deep guided filtering. Multimedia Systems **28**, 1449–1463 (2020). https://doi.org/10.1007/s00530-020-00706-0

[10] Zhao, Z., Bai, H., Zhang, J., *et al.*: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5906–5916 (2022). https://doi.org/10.1109/CVPR52729.2023.00572

[11] Zhao, Z., Bai, H., Zhang, J., *et al.*: Equivariant multi-modality image fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 25912–25921 (2023). https://doi.org/10.1109/CVPR52733.2024.02448

[12] Liu, R., Liu, Y., Wang, H., *et al.*: Saffusion: a saliency-aware frequency fusion network for multimodal medical image fusion. Biomedical Optics Express **16**, 2459–2481 (2025). https://doi.org/10.1364/BOE.555458

[13] Wang, M., Liu, Z., Li, K., et al.: Task-generalized adaptive cross-domain learning for multimodal image fusion. arXiv **abs/2508.15505** (2025). https://doi.org/10.48550/arXiv.2508.15505

[14] Zhu, X., Zhao, W., Liu, X., Wu, Z., Ren, S.: Multi-feature aggregation attention for efficient image super-resolution. The Visual Computer **42**, 85 (2026). https://doi.org/10.1007/s00371-025-04304-x

[15] Kui, X., Ji, Z., Zou, B., *et al.*: Iterative collaboration network guided by reconstruction prior for medical image super-resolution. IEEE Transactions on Computational Imaging **11**, 827–838 (2025). https://doi.org/10.1109/TCI.2025.3577340

[16] Chen, Z., Yang, L., Lai, J.-H., Xie, X.: Cunerf: Cube-based neural radiance field for zero-shot medical image arbitrary-scale super resolution. In: IEEE/CVF International Conference on Computer Vision, pp. 21128–21138 (2023). https://doi.org/10.1109/ICCV51070.2023.01937

[17] Fang, W., Tang, Y., Guo, H., *et al.*: Cycleinr: Cycle implicit neural representation for arbitrary-scale volumetric super-resolution of medical data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11631–11641 (2024). https://doi.org/10.1109/CVPR52733.2024.01105

[18] Xu, Y., Li, X., Jie, Y., *et al.*: Simultaneous tri-modal medical image fusion and super-resolution using conditional diffusion model. In: Proceedings of Medical Image Computing and Computer Assisted Intervention (2024). https://doi.org/10.1007/978-3-031-72104-5-61

[19] Stephen, J.K., Kapyela, G., Msoka, V.C., *et al.*: Medical image fusion under diffusion-inspired multiframe super-resolution framework. EURASIP Journal on Advances in Signal Processing **2025**, 45 (2025). https://doi.org/10.1186/s13634-025-01230-x

[20] Geng, P., Liu, S., Zhuang, S.: Multimodal medical image fusion by adaptive manifold filter. Computational and Mathematical Methods in Medicine **2015**, 1–9 (2015). https://doi.org/10.1155/2015/564748

[21] Yin, H.: Tensor sparse representation for 3-d medical image fusion using weighted average rule. IEEE Transactions on Biomedical Engineering **65**, 2622–2633 (2018). https://doi.org/10.1109/TBME.2018.2811243

[22] Li, Y., Zhao, J., Lv, Z., *et al.*: Multimodal medical supervised image fusion method by cnn. Frontiers in Neuroscience **15**, 638976 (2021). https://doi.org/10.3389/fnins.2021.638976

[23] Guo, K., Li, X.-f., Fan, T., *et al.*: Vanet: a medical image fusion model based on attention mechanism to assist disease diagnosis. BMC Bioinformatics **23**, 548 (2022). https://doi.org/10.1186/s12859-022-05072-4

[24] Gu, X., Xia, Y., Zhang, J.: Multimodal medical image fusion based on interval gradients and convolutional neural networks. BMC Medical Imaging **24**, 232 (2024). https://doi.org/10.1186/s12880-024-01418-x

[25] Safari, M., Fatemi, A., Archambault, L.: Medfusiongan: multimodal medical image fusion using an unsupervised deep generative adversarial network. BMC Medical Imaging **23**, 203 (2023). https://doi.org/10.1186/s12880-023-01160-w

[26] Zhao, Z., Bai, H., Zhu, Y., *et al.*: Ddfm: Denoising diffusion model for multi-modality image fusion. In: IEEE/CVF International Conference on Computer Vision, pp. 8048–8059 (2023). https://doi.org/10.1109/ICCV51070.2023.00742

[27] Qu, J., Huang, D., Shi, Y., Liu, J., Tang, W.: Diffusion-driven multi-modality medical image fusion. Medical & Biological Engineering & Computing **63**, 2105–2118 (2025). https://doi.org/10.1007/s11517-025-03300-6

[28] Wen, Y., Luo, B., Shi, W., *et al.*: Sat-net: Structure-aware transformer-based attention fusion network for low-quality retinal funduimages enhancement. IEEE Transactions on Multimedia **27**, 6198–6210 (2025). https://10.1109/TMM.2025.3565935

[29] Wang, W., He, J., Liu, H., *et al.*: Mdc-rht: Multi-modal medical image fusion via multi-dimensional dynamic convolution and residual hybrid transformer. Sensors **24**, 4056 (2024). https://doi.org/10.3390/s24134056

[30] Luo, F., Wu, D., Pino, L.R., *et al.*: A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. Scientific Reports **15**, 11657 (2025). https://doi.org/10.1038/s41598-025-93616-y

[31] Xie, X., Cui, Y., Tan, T., *et al.*: Fusionmamba: dynamic feature enhancement for multimodal image fusion with mamba. Visual Intelligence **2**, 37 (2024). https://doi.org/10.1007/s44267-024-00072-9

[32] Zhu, Y., Lv, L., Zhang, P., *et al.*: Interactive spatial-frequency fusion mamba for multi-modal image fusion. IEEE Transactions on Image Processing **35**, 2380–2392 (2026). https://doi.org/10.1109/TIP.2026.3662596

[33] Jie, Y., Zhou, F., Tan, H., *et al.*: Tri-modal medical image fusion based on adaptive energy choosing scheme and sparse representation. Measurement **204**, 112038 (2022). https://doi.org/10.1016/j.measurement.2022.112038

[34] Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing **19**(11), 2861–2873 (2010). https://doi.org/10.1109/TIP.2010.2050625

[35] Dai, T., Cai, J., Zhang, Y., *et al.*: Second-order attention network for single image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11057–11066 (2019). https://doi.org/10.1109/CVPR.2019.01132

[36] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2472–2481 (2018). https://doi.org/10.1109/CVPR.2018.00262

[37] Zhang, L., Dai, H., Sang, Y.: Med-srnet: Gan-based medical image super-resolution via high-resolution representation learning. Computational Intelligence and Neuroscience **2022**, 1744969 (2022). https://doi.org/10.1155/2022/1744969

[38] Ahmad, W., Ali, H., Shah, Z., *et al.*: A new generative adversarial network for medical images super resolution. Scientific Reports **12**, 9533 (2022). https://doi.org/10.1038/s41598-022-13658-4

[39] Zhang, Q., Hang, Y., Wu, F., *et al.*: Super-resolution of 3d medical images by generative adversarial networks with long and short-term memory and attention. Scientific Reports **15**, 20828 (2025). https://doi.org/10.1038/s41598-025-05783-7

[40] Lu, H., Mei, J., Qiu, Y., *et al.*: Information sparsity guided transformer for multi-modal medical image super-resolution. Expert Systems with Applications **261**, 125428 (2024). https://doi.org/10.1016/j.eswa.2024.125428

[41] Qin, J., Xiong, J., Liang, Z.: Cnn–transformer gated fusion network for medical image super-resolution. Scientific Reports **15**, 15338 (2025). https://doi.org/10.1038/s41598-025-00119-x

[42] Namatevs, I., Sudars, K., Nikulins, A., *et al.*: Denoising diffusion algorithm for single image inplaine super-resolution in cbct scans of the mandible. In: International Scientific Conference on Information Technology and Management Science of Riga Technical University, pp. 1–6 (2023). https://doi.org/10.1109/ITMS59786.2023.1031779

[43] Sadikov, A., Wren-Jarvis, J., Pan, X., Cai, L.T., Mukherjee, P.: Generalized diffusion mri denoising and super-resolution using swin transformers. arXiv preprint arXiv:2303.05686 (2023). https://doi.org/10.48550/arXiv.2303.05686

[44] Ji, Z., Zou, B., Kui, X., *et al.*: Global and local mamba network for multi-modality medical image super-resolution. Pattern Recognition **173**, 112888 (2025). https://doi.org/10.1016/j.patcog.2025.112888

[45] Horé, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: International Conference on Pattern Recognition, pp. 2366–2369 (2010). https://doi.org/10.1109/ICPR.2010.579

[46] Wang, W., Deng, L., Cao, Q., *et al.*: A general image fusion approach exploiting gradient transfer learning and fusion rule unfolding. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**, 6347–6364 (2026). https://doi.org/10.1109/TPAMI.2026.3655694

[47] Tang, L., Li, C., Ma, J.: Mask-difuser: A masked diffusion model for unified unsupervised image fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence **48**, 591–608 (2026). https://doi.org/10.1109/TPAMI.2025.3609323

[48] Ji, Z., Zou, B., Kui, X., *et al.*: Deform-mamba network for mri super-resolution. In: Medical Image Computing and Computer Assisted Intervention, pp. 242–252 (2024). https://doi.org/10.1007/978-3-031-72104-5-24