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.

MMIF-MESFusion：Coupling Modality-Specific Enhancement and Statistics- Guided Adaptive Fusion for Medical Image Fusion

Ming Gong

gongming2006@163.com

Nantong University

XinWang Wang

Wenzhou University

Xinlan Chen

Wuhan Polytechnic University

Wenze Kan

Jiamusi University

Article

Keywords: Multimodal medical image fusion, Spatial gated mamba, Modality-specific enhancement, Adaptive feature fusion, Statistical priors

Posted Date: August 6th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10017378/v1

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

Additional Declarations: No competing interests reported.

MMIF-MESFusion：Coupling Modality-Specific Enhancement and Statistics-Guided Adaptive Fusion for Medical Image Fusion

& *Ming Gong [a1], & Xinwang Wang [b], & Xinlan Chen [c], &Wenze Kan [d]

[a] *Affiliated Hospital of Nantong University ,NanTong ,JiangSu, China*

[b] *College of Chemistry and Materials Engineering, Wenzhou University, Wenzhou, China*

[c] *Wuhan Polytechnic University，Wuhan Polytechnic University School of Management, Hubei Wuhan ,China*

[d] *The First Affiliated Hospital of Jiamusi University, Jiamusi City, China*

Abstract: This paper addresses two challenges in multimodal medical image fusion: heterogeneous feature distribution discrepancies and adaptive cross-modal information integration. To tackle these, we propose MESFusion, a modality-specific enhancement and statistically guided adaptive fusion framework. The Modality-Aware Enhancement (MAE) module enhances complementary information from anatomical and functional modalities by using multi-scale gradient extraction and energy statistical modeling. For deep feature representation, the Spatial Gated Mamba (SGMam) module is introduced to capture long-range spatial dependencies and global contextual information. Additionally, the Statistical-Guided Adaptive Fusion (SDAF) module dynamically adjusts fusion behavior according to the first- and second-order statistical characteristics of feature channels. This enables effective integration of heterogeneous modal information. Extensive ablation studies and comparative experiments were conducted on three publicly available multimodal medical image fusion datasets, including CT-MRI, PET-MRI, and SPECT-MRI. n addition, an external PET-CT dataset (NAF-PROSTATE) was employed to evaluate the cross-dataset generalization capability of the proposed method. Experimental results demonstrate that MESFusion achieves competitive and well-balanced performance across multiple quantitative metrics while producing visually coherent fusion results with favorable structural preservation and information representation. These findings indicate the effectiveness and robustness of the proposed framework for multimodal medical image fusion. Our code will release at: https://github.com/bluemountain2025/MESFusion.

Keywords: Multimodal medical image fusion; Spatial gated mamba; Modality-specific enhancement; Adaptive feature fusion; Statistical priors

1. Introduction

With the rapid advancement of modern medical imaging technology, multimodal imaging has become indispensable in clinical practice. Frequently employed modalities, such as CT, MRI, PET, and SPECT, provide substantial assistance in medical diagnostics [1]. CT and MRI are examples of anatomical imaging. By virtue of their high spatial resolution, they can clearly delineate the morphology, localisation, and geometric characteristics of organs and lesions, yet they exhibit limited proficiency in reflecting tissue metabolic levels and early cellular functional changes [2]. Conversely, PET and SPECT serve as functional metabolic imaging modalities. These techniques possess high sensitivity in capturing biochemical metabolic processes and molecular-level anomalies within the body, offering unique advantages in early tumour screening. However, they are often constrained by low spatial resolution and a lack of clear anatomical background, posing challenges for the precise spatial localisation of lesions [3].

*Corresponding authors: Gongming (e-mail: gongming2006@163.com); & These authors contributed equally to this work and should be considered co-first authors

Nevertheless, due to the distinct physical imaging mechanisms underpinning these technologies, the pathological information they provide is characterised by significant heterogeneity and inherent limitations. Single-modal images often fail to deliver comprehensive and precise diagnostic information, prone to creating "information blind spots" in complex clinical decision-making [4]. Consequently, it is imperative to perform spatial registration and feature fusion of source images from different modalities to generate a synthetic image that simultaneously possesses high spatial-resolution anatomical details and rich functional metabolic information [5].

Deep learning-based image fusion methods are primarily classified into convolutional neural network (CNN), generative adversarial network (GAN), transformer approaches, and mamba approaches. CNN extract multi-scale spatial features via convolution kernels, effectively preserving local details and edge information within anatomical structures. For instance, the long-range correlation-guided network proposed by Zhou et al. [8] represents a transition from CNN-based local feature fusion to global correlation modelling. This work not only inherits the advantages of dual-encoder architectures in feature decoupling but also introduces long-range modelling concepts similar to Transformers. GAN leverages the adversarial game mechanism between the generator and discriminator to force the fused image to approximately match the distributions of high-quality reference images, thereby significantly improving texture details and visual clarity. Notably, Wang et al. [9] proposed the DiCyc (Deformation invariant cross-domain information fusion for medical image synthesis) model. By introducing a deformation-invariant module into a cycle-consistent adversarial network, this model effectively mitigates information distortion caused by anatomical misalignment during cross-modal synthesis. Transformer methods overcome the limitations of local receptive fields in traditional CNNs by computing correlations between pairs of feature patches in an image. For example, the CTGFSR (CNN-Transformer gated fusion network for medical image super-resolution) network proposed by Qin et al. [10] successfully improves the quality of medical image super-resolution reconstruction by integrating parallel dynamic convolution branches with residual Transformer branches and utilising gated attention mechanisms for feature consolidation. Mamba fusion methods represent a novel deep learning architecture based on a selective state space model (SSM). Its core innovation lies in the selective scan mechanism, which achieves adaptive compression and feature extraction of input sequences. Xie et al. [11] proposed dynamic feature enhancement for multimodal image fusion with mamba (FusionMamba), aiming to address the issues of global context loss and cross-modal information redundancy in multimodal image fusion. By demonstrating its superiority in achieving high-fidelity fusion with low computational overhead across multiple medical imaging datasets, this method provides an efficient end-to-end solution for high-resolution multimodal image fusion.

Despite the remarkable strides made by deep learning-based methods in medical image fusion, mainstream architectures still face numerous challenges. CNN are limited by local receptive fields, making it difficult to capture long-range dependencies, which results in suboptimal global contrast and brightness distribution in fused images; although GAN enhance texture generation capabilities, their inherent training instability and difficulty in convergence may introduce pseudo-lesion information; Transformers achieve global modeling via self-attention mechanisms, yet their quadratic computational complexity leads to massive GPU memory consumption; while the emerging Mamba architecture balances linear complexity with global receptive fields, its intrinsic 1D causal sequence modeling mechanism is ill-suited for the non-causal spatial structure of 2D/3D medical images, leading to impaired spatial continuity and insufficient cross-modal feature interaction. Furthermore, existing medical image fusion networks typically utilise identical encoders to process all inputs. However, MRI images are rich

in high-frequency anatomical details, whereas PET, SPECT, and CT images primarily exhibit low-frequency metabolic intensity distributions. Concurrently, most current fusion strategies primarily employ attention mechanisms to focus on local pixels while neglecting the holistic distributional properties of feature maps.

To address the aforementioned problems, this paper proposes a medical image fusion model coupling modality-specific enhancement with statistical-guided adaptive fusion. First, to overcome the significant discrepancies in feature distributions between MRI and functional images, we design a modality-aware enhancement module utilising gradient attention and intensity statistics, with distinct sub-modules for MRI and functional images that extract high-frequency textures and low-frequency energy information, respectively. Subsequently, in the feature extraction stage, we design an SGMam module to preserve local inductive biases and, combined with a multi-path scanning strategy, effectively reconstruct 2D spatial dependencies while maintaining linear computational complexity. Finally, to resolve the information redundancy and complex computational logic introduced by attention mechanisms, we propose an SDAF layer. By implementing a dynamic weighting mechanism for the complementary integration of cross-modal features, this approach significantly improves the quality of multimodal medical image fusion. The main contributions of this paper are summarised as follows:

1） We propose MESFusion, a lightweight and end-to-end multimodal medical image fusion framework that is specifically designed for heterogeneous medical modalities. Rather than stacking generic modules, MESFusion follows a problem-driven encoder–fusion–decoder paradigm that jointly addresses three coupled challenges in heterogeneous medical fusion: shallow representation discrepancy across modalities, insufficient 2D long-range dependency modeling in efficient backbones, and lightweight yet distribution-aware cross-modal aggregation.

2） To reduce the representation gap between heterogeneous modalities at the shallow stage, we develop a modality-aware enhancement (MAE) module that performs differentiated feature calibration for anatomical and functional images according to their distinct imaging characteristics. Unlike homogeneous shared preprocessing, MAE enables modality-specific enhancement before deep interaction, thereby improving the preservation of anatomical structures and functional responses in early representations.

3） To overcome the limitations of directly applying standard Mamba-style sequence modeling to 2D medical images, we design a lightweight spatial-gated Mamba backbone, termed SGMam, which jointly captures local spatial continuity and long-range dependencies with low computational overhead. By integrating background-aware spatial gating and multi-path selective scanning, SGMam improves global contextual modeling while preserving fine-grained anatomical continuity, making it more suitable for medical image fusion than a direct adoption of conventional Mamba variants.

4） We propose a statistically guided adaptive fusion (SDAF) layer for lightweight cross-modal feature aggregation. By jointly exploiting channel-wise mean and standard deviation as complementary statistical priors, SDAF performs distribution-aware adaptive fusion and dynamically balances structural and functional information, providing an efficient alternative to computationally intensive attention-based fusion strategies. Extensive ablation results further demonstrate that the performance gains arise from the complementary synergy among MAE, SGMam, and SDAF, rather than from isolated module stacking.

The remainder of this paper is organised as follows. Section Ⅱ reviews related work on medical image fusion in recent years. Section Ⅲ elaborates on the architecture of the proposed MESFusion network. Section Ⅳ presents the experimental results and analysis, including ablation studies on the

network architecture and hyperparameters of the loss function, as well as comparative experiments against seven state-of-the-art image fusion algorithms. Finally, Section Ⅴ concludes the paper with a discussion, summary, and an outlook on future research directions.

2. Related work

2.1 Traditional image fusion methods

Traditional medical image fusion methodologies are primarily categorised into spatial-domain and transform-domain approaches. Spatial-domain-based methods primarily include independent component analysis (ICA), principal component analysis (PCA), and weighted averaging [12]. Conversely, transform domain-based methods typically operate by converting input images from the spatial domain to the frequency domain via the Fourier Transform, applying specific fusion algorithms to the transformed coefficients, and subsequently reconstructing the final fused image via the Inverse Fourier Transform. Prominent algorithms within the transform domain include pyramid transforms, wavelet transforms, and multi-scale geometric analysis (MGA) [13].

2.2 CNN image fusion methods

CNN demonstrate robust efficacy in processing spatial and structural information within local neighbourhoods of input images. A typical CNN architecture comprises a sequence of convolutional, pooling, and fully connected layers, where convolutional and pooling layers extract features from source images, and fully connected layers map these features to the final output. For instance, Zhao et al. [14] proposed a denoising diffusion fusion model (DDFM), which formulates the fusion task as a conditional generation problem conditioned on source images. Similarly, Guan et al. [15] introduced an edge-enhancement multi-scale feature fusion network (EMFF-Net), which employs a specialised multi-scale feature extraction module to capture holistic anatomical structures and local textures, respectively. Subsequently, the fusion network integrates these two distinct sets of information, thereby enhancing the CNN's capability to capture both global structures and local details simultaneously..

2.3 GAN fusion methods

Recent advancements in GANs have incorporated attention mechanisms to guide generators toward regions of interest or introduced dual—or even triple—discriminators to balance information across different modalities. Furthermore, to mitigate distortion in generated images, researchers have begun to impose strict edge and structural constraints within the loss functions. For example, to address edge blurring in medical images, Chao et al. [15] proposed constrained generative adversarial networks (CGAN), which embed an explicit edge-constraint module into the GAN framework. By utilising edge information extracted via operators such as the Sobel operator as prior knowledge, it forces the generator to produce fused images with distinct tissue contours, effectively resolving the boundary bleeding issue in MRI fusion. Additionally, Li et al. [16] proposed a hybrid generative adversarial network based on a mixed attention fusion (HGAN), in which spatial and channel attention mechanisms are embedded in both the generator and the discriminator. Rather than merely relying on adversarial learning to fool the discriminator, this approach utilises attention maps to guide the network to focus on highlighted information while suppressing irrelevant noise, thereby achieving more precise injection of functional information.

2.4 Transformer fusion methods

To overcome the limitations of CNN receptive fields, researchers have introduced transformers to capture long-range dependencies. these networks typically operate by tokenising images into patches, utilising self-attention mechanisms to compute weights between different modalities, globally integrating

features, and reconstructing the fused result based on these weights. for instance, tang et al. [18] proposed multimodal medical image fusion via multiscale adaptive transformer (MATR). MATR designs a multi-scale adaptive transformer module that maps image features into a common latent space and subsequently utilises an adaptive attention mechanism to automatically adjust the scope of attention based on the complexity of image content, thereby assigning varying degrees of focus to different regions. ma et al. [19] introduced cross-domain long-range learning for general image fusion via the swintransformer (SwinFusion), which employs a shifted-window mechanism to facilitate information exchange across windows while maintaining computational efficiency within local windows. furthermore, a cross-domain attention module is designed to enable deep interaction of features from different modalities as early as the feature extraction stage.

2.5 Mamba fusion methods

To address the high memory footprint and slow inference speed associated with Transformers, researchers have proposed the Mamba network for deep learning. Mamba's linear computational complexity makes inference significantly more efficient when processing high-resolution medical images. For example, Xie et al. [11] proposed FusionMamba, which incorporates a dynamic interaction Mamba module. Unlike the global attention in Transformers, FusionMamba adaptively selects important feature channels for scanning and fusion, maintaining soft-tissue contrast in MRI while achieving processing speeds several times faster than comparable Transformer methods. Li et al. [20] proposed MambaDFuse, which utilises Mamba blocks to process low-frequency and high-frequency information separately, yielding fused images that are visually clear with minimal artefacts. In this paper, we integrate the Mamba network with a spatial gating mechanism to propose the SGMam module for deep feature extraction, providing reliable information for the subsequent fusion of heterogeneous modalities.

3. Methodology

In this section, we elaborate on the architectural details of the proposed MMIF-MESFusion framework. Specifically, Section A presents the overall structure of the fusion network. Section B details the MAE module. Section C introduces the SGMam feature extraction module. Section D elucidates the proposed SDAF mechanism. Section E outlines the decoder's design. Finally, Section F formulates the loss functions utilised for model training.

3.1. Overall Architecture

The MESFusion framework employs an end-to-end backbone network comprising four principal stages: image feature enhancement, deep feature extraction, information fusion, and image reconstruction. To address the distinct imaging mechanisms across heterogeneous modalities, the network first performs specific pre-enhancement steps. It utilises multi-scale gradient operators to accentuate the anatomical structural features of MRI, and energy-aware attention mechanisms to highlight metabolic intensity features of functional imagery. Next, the SGMam module serves as the backbone feature extractor, incorporating embedded spatial gating units and a 2D selective-scanning mechanism to efficiently capture long-range dependencies and global context while maintaining linear computational complexity. The extracted deep features are then processed by the SDAF layer, which dynamically calibrates channel weights based on the global mean and standard deviation of the feature maps, thereby achieving optimal integration of cross-modal complementary information. Finally, the fused feature maps are fed into a multi-layer convolutional decoder to reconstruct the final fused image, characterised by both sharp anatomical contours and precise functional information.

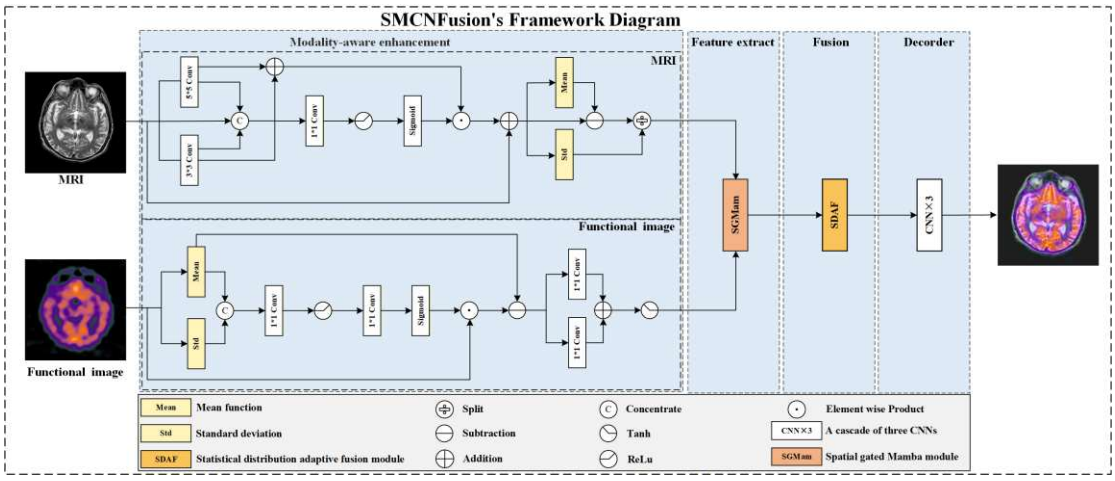


Fig.1. The overall framework of the proposed practical medical image fusion network

3.2 Modality-aware enhancement module

Current medical image fusion frameworks typically employ a unified module for feature enhancement across different inputs. However, this approach overlooks the fact that distinct medical imaging modalities harbour disparate information. For instance, MRI images predominantly provide rich anatomical structures and soft-tissue textures, whereas functional images primarily reflect tissue density, metabolic activity, or functional activity. Consequently, we devise a differentiated feature enhancement strategy to maximise the representation of characteristic information for each specific modality. The detailed structure of the MAE module is illustrated in Fig. 1.

To capture structural information across multiple scales, the MAE module employs a multi-scale convolutional strategy to extract high-frequency gradient features from MRI images. Specifically, we utilise 3×3 and 5×5 grouped convolution operators as learnable gradient filters to extract fine-grained and coarse-grained gradient responses from the input feature x, respectively. By computing the absolute values of these responses and fusing them, we obtain the output result G_1. This process can be formulated as:

$$G_1 = \left|\mathrm{Conv}_1(x)\right| + \left|\mathrm{Conv}_2(x)\right| \tag{1}$$

Here, Conv1 and Conv2 denote convolution operations with kernel sizes of 3×3 and 5×5, respectively. This multi-scale architectural design facilitates the simultaneous capture of subtle texture variations and prominent structural contours. To mitigate the introduction of redundant background noise, we construct a structure-aware attention mechanism to modulate the injection of gradient information. Specifically, this mechanism concatenates the original feature x with the multi-scale gradient map G1 along the channel dimension, serving as contextual input for the attention network. This process can be formulated as:

$$G_2 = \mathrm{ConCat}\left(x, \mathrm{Conv}_1(x), \mathrm{Conv}_2(x)\right) \tag{2}$$

$$G_3 = \mathrm{Sigmoid}\left(\mathrm{ReLu}\left(\mathrm{Conv}_3(G_2)\right)\right) \tag{3}$$

Where ConCat denotes the channel-wise concatenation operation, ReLu represents the rectified linear unit activation function, and Conv3 refers to a convolutional layer with a kernel size of 1×1.

Subsequently, the derived attention map G3 is utilised to modulate the gradient features G1. The resulting weighted features are then injected into the original input via a residual connection, yielding

the final structure-enhanced feature representation Gj. This process can be formulated as:

$$G_{\mathrm{j}} = G_3 \odot G_1 + x \tag{4}$$

Here, ⊙ denotes the element-wise product. Given the significant contrast variations inherent in MRI images across diverse acquisition protocols, direct fusion is prone to inducing mismatches in intensity distribution. Consequently, we introduce an adaptive contrast normalisation strategy. This strategy computes the mean μ and standard deviation σ of the input features along the spatial dimensions to standardise the feature representations. Subsequently, a dynamic affine transformation is performed via convolutional layers to generate the final output G_{m}. This process can be formulated as:

$$\mu = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} G_{\mathrm{j}}(i, j) \tag{5}$$

$$\sigma = \sqrt{\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(G_{\mathrm{j}}(i, j) - \mu \right)^2} \tag{6}$$

$$G_{\mathrm{m}} = \mathrm{Conv}_1 \left(\frac{x - \mu}{\sigma} \right) \tag{7}$$

Where H and W denote the height and width of the input features, respectively, and (i, j) represents the spatial coordinates of the pixels.

Distinct from the rich anatomical textures inherent in MRI images, the pixel intensity values in functional imagery typically correspond directly to specific physical quantities or physiological indicators. To accurately preserve this intensity information and highlight high-metabolic regions during feature extraction, we devise a feature enhancement strategy distinct from that used in MRI.

To capture the intensity distribution characteristics from a global perspective, we first perform channel-based statistical modelling on the input feature y. Specifically, we compute the global mean μ_{y} and standard deviation σ_{y} of the feature maps across spatial dimensions, serving as statistical descriptors of the image's overall energy level and dispersion, respectively. Subsequently, these statistics are concatenated along the channel dimension to construct the statistical feature vector S. To adaptively recalibrate the importance of each channel based on these statistical features, we design a lightweight statistical encoder composed of two convolutional layers. Finally, a Sigmoid function is applied to generate the channel-level energy-aware attention weights S1. This process can be formulated as:

$$S = \mathrm{ConCat}\left(\mu_{\mathrm{y}}, \sigma_{\mathrm{y}} \right) \tag{8}$$

$$S_1 = \mathrm{Sigmoid}\left(\mathrm{Conv}_3 \left(\mathrm{ReLu}\left(\mathrm{Conv}_3 \left(S \right) \right) \right) \right) \tag{9}$$

Subsequently, an element-wise product is performed between the energy-aware attention map S_1 and the input features. This operation adaptively accentuates feature channels that are rich in functional information. However, because medical functional imagery typically exhibits a broad dynamic range, direct manipulation is prone to numerical instability. To mitigate this, we incorporate a dynamic range rescaling mechanism. Specifically, we zero-centre the feature distribution by subtracting the global mean μ_{y} from the enhanced features. This process can be formulated as:

$$S_2 = x \odot S_1 \tag{10}$$

$$S_3 = S_2 - \mu_{\mathrm{y}} \tag{11}$$

Where S_2 denotes the result of the element-wise product, and S_3 represents the zero-centered feature maps.

To preserve relative intensity discrepancies while simultaneously suppressing outliers, we employ two learnable convolutional layers to perform non-linear mapping and shifting on the centred features. Specifically, we use the hyperbolic tangent (Tanh) function to apply nonlinearity to the scaling component, thereby improving the robustness of the features against numerical fluctuations.

$$S_{\mathrm{F}} = \mathrm{Tanh}\left(\mathrm{Conv}_{3\alpha}\left(S_3\right) + \mathrm{Conv}_{3\beta}\left(S_3\right)\right) \tag{12}$$

Here, Tanh denotes the Hyperbolic Tangent function, while $\mathrm{Conv}_{3\alpha}$ and $\mathrm{Conv}_{3\beta}$ refer to two distinct 1×1 convolutional layers with independent parameters.

3.3 Spatial gated mamba module

Efficient and precise feature extraction is imperative for multimodal medical image fusion; however, conventional approaches face significant challenges. Common weight-sharing extractors often overlook inherent modality discrepancies and treat pixels indiscriminately, failing to account for the substantial background redundancy and imaging noise typical in medical data. This indiscriminate processing leads to attentional dispersion and suboptimal salient-region representation. To address these limitations, we propose the SGMam module, a novel architecture that synergises linear global modelling with local feature extraction within a unified framework. As illustrated in Fig. 2, SGMam is designed to capture distinct modality-specific information while mitigating background interference, thereby ensuring robust feature representation and precise fusion.

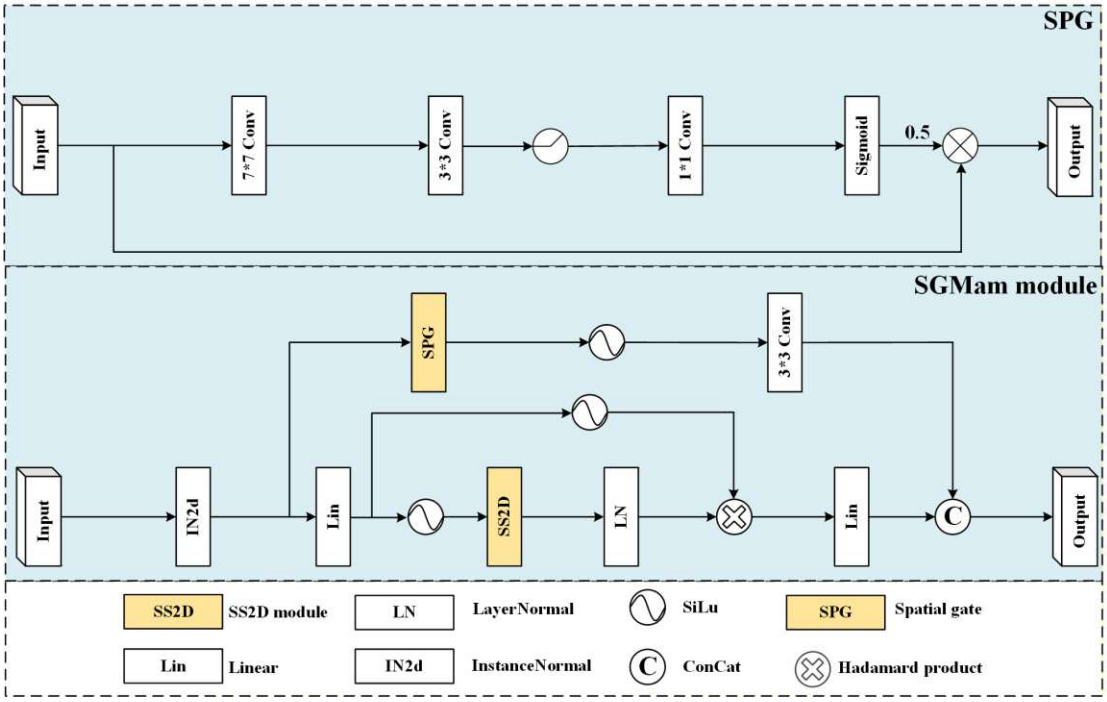


Fig. 2. Illustration of the SGMam module architecture

To address this, the SPG module exploits local spatial context information to generate an attention mask, thereby suppressing trivial information while preserving critical structural details. Initially, the input features undergo instance normalisation to standardise the feature distribution. Subsequently, the features are fed into two distinct branches. Branch 1, the SPG module, performs saliency calibration on the input information. This module utilises a 7×7 depth-wise convolution to capture local spatial context; the enlarged kernel size expands the receptive field, enabling the module to encompass broader local spatial neighbourhoods while maintaining low computational overhead. Following this, two lightweight convolutional layers are employed to first reduce and then expand the channel dimensions, aiming to

trade minimal computational cost for enhanced non-linear expressive capability and compelling the model to distil the most core anatomical correlations within the compressed feature space, thereby generating a stable and robust spatial weight map w. This process can be formulated as:

$$w = \mathrm{Sigmoid}\left(\mathrm{Conv}_3\left(\mathrm{ReLu}\left(\mathrm{Conv}_1\left(\mathrm{Conv}_4\left(\mathrm{IN2d}\left(x\right)\right)\right)\right)\right)\right) \tag{13}$$

Here, x denotes the input information, and Conv4 represents the 7×7 convolution operation.

Subsequently, the generated weights are reapplied to the original features; this residual connection effectively mitigates the vanishing gradient problem often encountered in deep networks during initial training and ensures the structural integrity of medical anatomy. This process is formulated as:

$$P_1 = x \times \left(0.5 + 0.5 \times w\right) \tag{14}$$

Subsequently, P_1 undergoes instance normalisation to standardise the feature distribution, and the resulting representation is fed into two distinct branches. Upon enhancement by the SPG module, the feature representations are locally refined via a 3×3 standard convolutional layer; this operation is explicitly designed to capture fine-grained textures in medical imagery, thereby ensuring edge fidelity in the fused image. This process can be formulated as:

$$P_s = \mathrm{Conv}_1\left(\mathrm{SiLu}\left(P_1\right)\right) \tag{15}$$

Here, P_s is the output from the first branch. In the second branch, we add an improved 2D SSM. First, we double the input feature dimensions to prepare for non-linear processing. Next, the features are split into a main branch and a gating branch. The main branch uses a selective scan function that runs in parallel along four directions. By using simplified State Space Equations, this branch captures long-range anatomical patterns with low computation:

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t \tag{16}$$

$$y_t = Cx_t + Dh_{t-1} \tag{17}$$

Here, x_t denotes the input signal at time step t, corresponding to the t-th pixel in the scanning sequence within the context of image processing; h_t signifies the hidden state or state variable of the system at time instance t; and $\bar{A}$ represents the discretized state transition matrix, $\bar{B}$ means input matrix, $\bar{C}$ represents output matrix, and $\bar{D}$ is the direct feed-through matrix, respectively. y_t denotes the output signal at time step t, corresponding to the t -th pixel of the scanning sequence.

The gating branch interacts with the main branch's result via a Hadamard product, followed by an activation function, to yield the final output. This process can be formulated as:

$$P_m = \mathrm{SiLu}\left(\mathrm{Lin}\left(x_1\right)\right) \otimes \mathrm{LN}\left(\mathrm{SS2D}\left(\mathrm{SiLu}\left(\mathrm{Lin}\left(x_1\right)\right)\right)\right) \tag{18}$$

Here, SS2D denotes the 2D selective scan, $\otimes$ represents the hadamard product, LN signifies layer normalization, and Lin refers to a linear layer. x_1 corresponds to the resulting features subsequent to the instance normalization (IN2d). The hidden state dimension is set to d_{state}=16, and the expansion ratio is set to 2 (expand = 2). The time-step rank is adaptively set as 2 ($\mathrm{d_t}$=⌈$\mathrm{d_{model}}$/16⌉). Finally, the outputs derived from the two branches are aggregated via concatenation to yield the final output P_o of the SGMam module.

$$P_o = \mathrm{ConCat}\left(\mathrm{Lin}\left(P_m\right), P_s\right) \tag{19}$$

Compared with the standard VMamba block, the proposed SGMam introduces two lightweight modifications: an SPG branch before sequence modeling to suppress background redundancy and

enhance critical structural responses, and a parallel local-refinement branch to complement local texture modeling and preserve fine-grained anatomical continuity. Specifically, the SPG branch employs a 7×7 depth-wise convolution followed by two point-wise convolutions for channel compression and restoration, thereby generating a spatial weight map for adaptive feature recalibration. The outputs of the local-refinement and SS2D branches are concatenated along the channel dimension and projected by a linear layer to form the final output. This design preserves the linear-complexity advantage of state-space modeling while introducing only lightweight convolutional overhead. The lightweight property of SGMam is further validated in the experimental section through comparisons of Flops and parameters with other representative methods.

3.4 SDAF module

In current multimodal medical image fusion paradigms, rudimentary operations such as feature concatenation or element-wise addition frequently overlook the disparate importance of multi-modal features across spatial and channel dimensions, thereby predisposing the system to the submersion of critical information; conversely, the deployment of attention mechanisms inevitably escalates the model's computational complexity. To address these challenges, we propose the SDAF layer. This module achieves dynamic weight calibration and optimised integration of cross-modal information by explicitly modelling the global first-order and second-order statistics of the feature maps, as illustrated in the detailed architecture shown in Fig. 3.

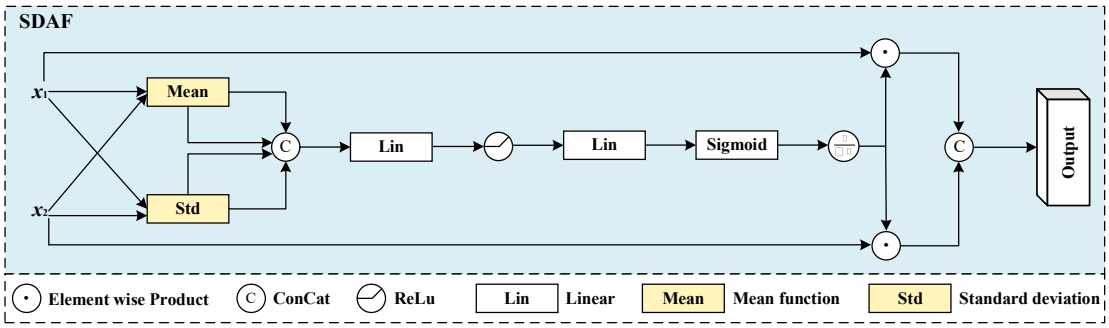


Fig. 3. Schematic diagram of the SDAF module architecture

To perceive the energy distribution and contrast characteristics inherent in heterogeneous modal feature streams, the SDAF module first extracts global contextual statistical information from the input features x_1(B, C_1, H, W) and $x2$(B, C_2, H, W). For each channel, the mean μ_f and standard deviation σ_f are computed, where the mean signifies the global energy intensity of the modal features, while the standard deviation reflects the dispersion and richness of the textures. By concatenating channel-wise statistics from both feature streams, a global statistical feature descriptor $S = [\mu_1, \mu_2, \sigma_1, \sigma_2]$ is constructed. To derive optimal fusion weights from this complex statistical descriptor, we design a lightweight bottleneck mapping network. This network comprises two fully connected layers with an embedded ReLU activation function to enhance non-linear modelling capabilities:

$$w = \mathrm{Sigmoid}\Big(\mathrm{Lin}\Big(\mathrm{ReLu}\big(\mathrm{Lin}\big(S\big)\big)\Big)\Big) \tag{20}$$

To ensure the stability of the total information magnitude during the fusion process and to avoid potential numerical overflow caused by the disproportionate dominance of specific modal weights, we employ a normalisation mechanism. Specifically, for the generated raw weight vector w, the final channel-wise weights ω are computed as follows:

$$\omega = \frac{w}{\sum w + \varepsilon} \tag{21}$$

Here, ε represents a small constant introduced to ensure numerical stability. Upon obtaining the adaptive weights, we decouple the weight vector into distinct components, ω_1(B, C_1) and ω_2(B, C_2), corresponding to the two respective input streams. Subsequently, channel modulation is performed on the original feature maps via element-wise multiplication, thereby achieving the simultaneous accentuation of critical features and suppression of redundant information:

$$d_1 = x_1 \square \ \omega_1 \tag{22}$$

$$d_2 = x_2 \square \ \omega_2 \tag{23}$$

Here, x_1 denotes the first input modality, and x_2 represents the second input modality. Finally, the modulated feature maps undergo a concatenation operation along the channel dimension to yield the final fused feature representation:

$$d_f = [d_1; d_2] \tag{24}$$

The architectural design of the SDAF module balances computational efficiency with fusion performance. By incorporating standard deviation statistics, the model shows superior sensitivity in capturing fine-textured edges within MRI scans and high-intensity functional regions in functional images. It significantly surpasses methods that rely only on mean-based statistics. This adaptive weight allocation mechanism allows MESFusion to dynamically balance anatomical structures and functional information under different imaging conditions. The approach depends on the specific content of the input imagery, thus generating fused images with enhanced information representation capability.

3.5 Decoder

After deep feature extraction and cross-modal adaptive fusion, the model is tasked with mapping abstract features from the high-dimensional latent space back to the original pixel space. To this end, we design an image reconstruction decoder characterised by a hierarchical optimisation structure, aiming to restore spatial details and ensure the radiometric fidelity of the generated results. The detailed architecture is illustrated in Fig. 4.

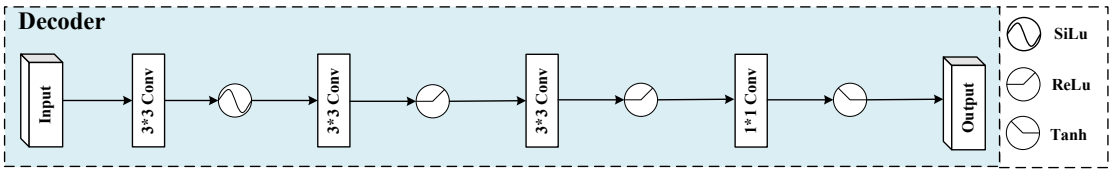


Fig. 4. Detailed architecture of the decoder

The decoder adopts a cascaded convolutional architecture comprising four functional layers, designed to refine features and restore spatial information by progressively compressing channel dimensions. Initially, the fused feature maps pass through a 3×3 convolutional layer, where the channel count is reduced from the input dimension to a preset 32. This layer is coupled with the SiLU activation function, which leverages its smooth, non-monotonic properties to better preserve subtle gradient information than the traditional ReLU. Subsequently, two distinct 3×3 convolutional layers with ReLU activation functions are employed to reduce the channel dimension to 16 and 8, respectively. This design enhances the textural representation of anatomical structures by augmenting the network's non-linear depth. The final 1×1 convolutional layer reduces the channel count to 1, resulting in a single-channel grayscale image. This layer utilises the Tanh activation function to strictly constrain the output values within the [-1, 1] interval, thereby maintaining consistency with the distribution of the input images.

3.6 Loss Function

To ensure that the fused results simultaneously preserve the fine-grained textural edges inherent in

MRI and highlight salient functional regions within functional imagery, we construct a multi-dimensional composite loss function. This objective function comprises three distinct components: intensity loss (L_{in}), gradient loss (L_{gd}), and structural similarity loss (L_{sm}). The total loss is mathematically formulated as:

$$L_{\text{fuse}} = aL_{\text{in}} + bL_{\text{gd}} + cL_{\text{sm}} \tag{25}$$

Here, a, b, and c represent the hyperparameters of the loss function.

3.6.1 Intensity loss

The intensity loss is designed to constrain the pixel intensity distribution of the fused image. We integrate strategies focused on mean background preservation and salient target enhancement:

$$L_{\text{in}} = \left\| I_{\text{f}} - \frac{I_a - I_b}{2} \right\|_2^2 + 0.5 \left\| I_{\text{f}} - \max\left(I_{\text{a}} - I_{\text{b}}\right) \right\|_1 \tag{26}$$

Here, I_{f} denotes the fused result, while I_{a} and I_{b} represent the MRI image and the functional image, respectively. The first term of the equation encourages the fused image to maintain an overall intensity balance. The second term, by computing the element-wise maximum of the MRI and functional images, compels the model to focus on regions of interest, thereby effectively highlighting the targets.

3.6.2. Gradient loss

To extract fine-grained textures and edge details from both source images, we define a loss function based on the pixel-level maximum gradient response:

$$L_{\text{gd}} = \left\| \nabla I_{\text{f}} - \max\left(\left|\nabla I_{\text{a}}\right|, \left|\nabla I_{\text{b}}\right|\right) \right\|_2^2 \tag{27}$$

Where ∇ denotes the gradient operator. Distinct from traditional sum-of-squares methods, we employ a maximum selection strategy to retain the strongest edge features from both modalities as the target, which effectively mitigates artifact generation.

3.6.3. Structural similarity loss

To ensure that the fused image maintains perceptual consistency with the source images, we introduce the MS-SSIM loss:

$$\text{MS-SSIM}\left(I_{\text{a}}, I_{\text{b}}\right) = \left[l_{\text{M}}\left(I_{\text{a}}, I_{\text{b}}\right)\right]^{\alpha_{\text{M}}} \prod_{j=1}^{\text{M}} \left[c_j\left(I_{\text{a}}, I_{\text{b}}\right)\right]^{\beta_j} \left[s_j\left(I_{\text{a}}, I_{\text{b}}\right)\right]^{\gamma_j} \tag{28}$$

$$L_{\text{sm}} = 1 - \frac{1}{2}\left(\text{MS-SSIM}\left(I_{\text{f}} - I_{\text{a}}\right) + \text{MS-SSIM}\left(I_{\text{f}} - I_{\text{b}}\right)\right) \tag{29}$$

Here, M denotes the total number of scales. The algorithm performs iterative downsampling on the original image M-1 times, generating a cascade of images with progressively diminishing resolutions. j represents the index of the current scale. α_{M} signifies the weight governing the luminance information at the lowest scale, while β_j and γ_j represent the sensitivity parameters controlling the contrast and structural information at the j-th scale, respectively.

4.Experimental validation

4.1. Experimental details

To comprehensively validate the effectiveness and superiority of MESFusion, we conducted both comparative experiments and ablation studies across three publicly available multimodal medical image fusion datasets: SPECT-MRI, CT-MRI, and PET-MRI [21]. The dataset partitions were defined as follows: the CT-MRI dataset contains 160 training pairs and 24 testing pairs, the SPECT-MRI dataset contains 300 training pairs and 57 testing pairs, and the PET-MRI dataset contains 200 training pairs and

69 testing pairs. Furthermore, to further assess the generalization capability of the proposed method, an additional cross-dataset generalization experiment was performed on the NAF-PROSTATE dataset [22]. This dataset comprises F-18 PET/CT images from which 131,882 aligned slice pairs were extracted. Without any additional training or fine-tuning, the trained model was directly applied to this dataset for testing, thereby evaluating its generalization performance under different data distributions and clinical scenarios. To ensure a fair comparison, all competing methods were evaluated under the same testing protocol and on the same dataset partitions. In all tables, the best results are highlighted in red, while the second-best results are highlighted in blue.

To enhance the statistical reliability of the experimental results, we report not only the mean values of all quantitative metrics but also their corresponding 95% confidence intervals (CIs). Specifically, the metric value of each test sample is regarded as an independent observation. For each metric, the mean and standard deviation are first calculated across all test samples, and the 95% confidence interval for the mean is then estimated using the t-distribution. The corresponding formula is given as follows:

$$\mathrm{CI}_{95\%} = \bar{x} \pm t_{0.975,n-1} \cdot \frac{s}{\sqrt{n}} \tag{30}$$

where $\bar{x}$ denotes the sample mean, s denotes the sample standard deviation, n is the number of test samples, $t_{0.975,n-1}$ is the critical value of the t-distribution with n-1 degrees of freedom at the 95% confidence level.

During training, all source images were first converted to grayscale and uniformly resized to 224 × 224. To ensure consistent input dimensions, images smaller than the target size were resized using bicubic interpolation. In addition, all input intensities were normalized to the range [-1, 1] to stabilize optimization. During inference, the network generated fused images in a single-channel intensity space, which were then converted to RGB format only for visualization and qualitative comparison.

All experiments were conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (24 GB VRAM, a graphics processing unit for high-speed computations) and an Intel(R) Xeon(R) Platinum 8481C CPU (central processing unit), running a Python 3.10.15-based software environment. Model training proceeded for 200 epochs (full passes through the dataset) with a batch size of 16 (samples per gradient update) using the AdamW optimizer (a variant of Adam with weight decay for regularization). An initial learning rate of 1×10^{-5} (which controls the magnitude of weight adjustments) was used, gradually reduced using a StepLR scheduler (step-wise learning rate reduction) to improve convergence stability. Unless otherwise specified, loss function hyperparameters (a, b, and c, which determine specific aspects of the loss computation) were set to $a = 1$, $b = 2$, and $c = 3$.

Since multimodal image fusion lacks standard ground truth references and operates as an unsupervised learning task, we primarily evaluated fusion quality through a combination of subjective observation and objective metric comparison [23]. Subjective observation evaluates clarity, detail richness, and contrast based on human visual perception, while objective metric comparison measures fusion quality using quantitative values from fused and source images, providing concrete results to ensure a more objective and accurate evaluation. In terms of objective assessment, this study selected six metrics where red in tables indicates the best performance and blue denotes the second-best: VIF (Visual Information Fidelity): a perceptual metric that quantifies how much of the original visual information remains in the processed image, with higher values meaning more information is preserved. The correlation coefficient (CC) describes the degree of linear relationship (how closely two variables change together), with values closer to 1 indicating more similar spectral features. The edge-based fusion quality index (QAB/F) measures how well a fusion method preserves important edges and structures from the

source images; higher scores indicate better preservation of these features. Information entropy (EN) reflects the richness or variety of information in the image; higher values represent more detailed content. Peak signal-to-noise ratio (PSNR) measures image quality by comparing the signal strength to the noise, with higher PSNR indicating less distortion. The Multi-Scale Structural Similarity Index (MS-SSIM) evaluates similarity in luminance, contrast, and structure across different image resolutions and more closely aligns with human visual judgment than standard SSIM.

4.2. Ablation studies

To verify the effectiveness of the MESFusion network architecture, we conducted ablation experiments focusing on both the network structure and the hyperparameters of the loss function.

4.2.1. Ablation study on network architecture

To further validate the effectiveness of each component in the proposed MESFusion architecture, we conducted a comprehensive ablation study by systematically removing or replacing key modules. The first three settings correspond to single-module ablations: the MAE module is removed in the first setting, the SGMam module is removed in the second setting, and the SDAF module is replaced by a plain concatenation operation in the third setting. These experiments are designed to evaluate the individual contribution of each core module.

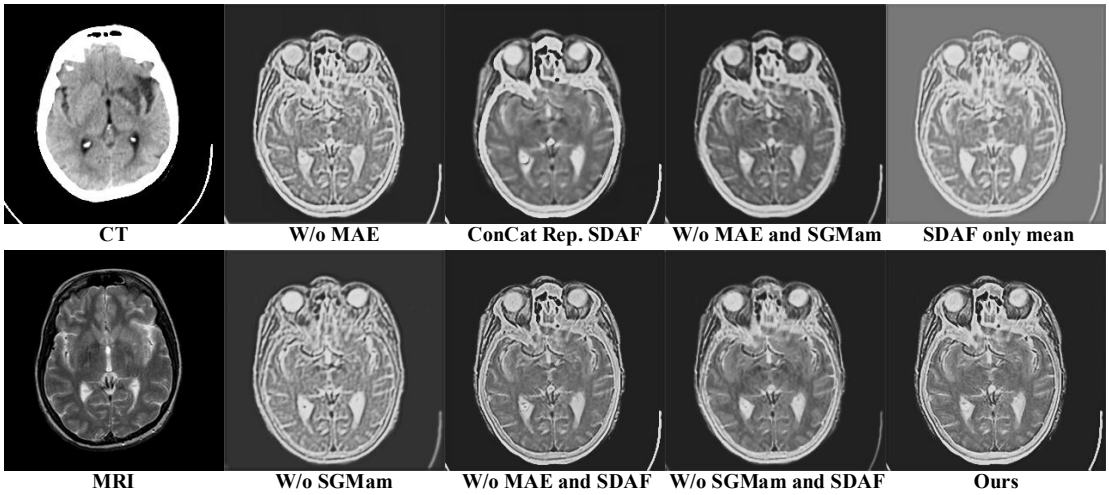


Fig.5. Subjective visualization of ablation studies on the CT-MRI dataset

The fourth to sixth settings correspond to combined ablations, which further investigate the collaborative effects among modules. In the fourth setting, both MAE and SDAF are removed, leaving only the Mamba-based backbone for feature extraction and fusion, in order to assess whether the backbone alone is sufficient for high-quality fusion. In the fifth setting, both MAE and SGMam are removed, and the Mamba-based extractor is replaced with a lightweight convolutional backbone to evaluate fusion performance without modality-specific enhancement or long-range dependency modeling. In the sixth setting, only the shallow MAE module is retained, while the SGMam module is replaced with a convolutional backbone and the SDAF module is replaced with plain concatenation, thereby assessing whether shallow modality-specific enhancement alone can provide sufficient performance gain. In the seventh setting, an internal ablation is performed on the SDAF module, where only the channel-wise mean is used for adaptive weight generation, and the standard deviation term is removed. This experiment is designed to verify whether first- and second-order statistical descriptors are complementary and whether using both the mean and standard deviation is more effective than using the mean alone. The eighth setting corresponds to the full MESFusion model, which includes all proposed

components. Figs.5, 6, and 7 present the qualitative ablation results on the CT-MRI, PET-MRI, and SPECT-MRI datasets. Overall, the visual comparisons show that different components of MESFusion contribute to complementary aspects of fusion quality, including preservation of anatomical detail, expression of functional information, and cross-modal consistency.

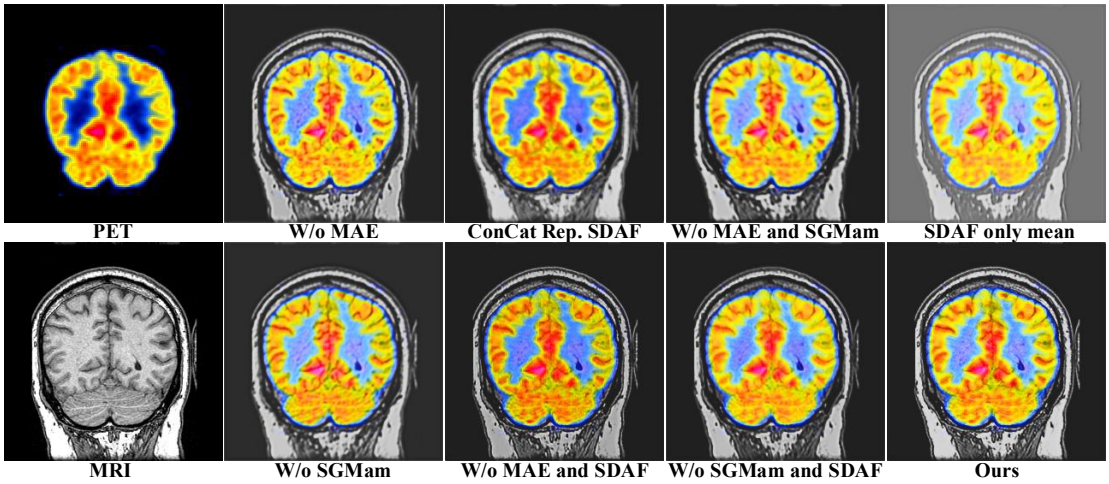


Fig.6. Subjective visualization of ablation studies on the PET-MRI dataset

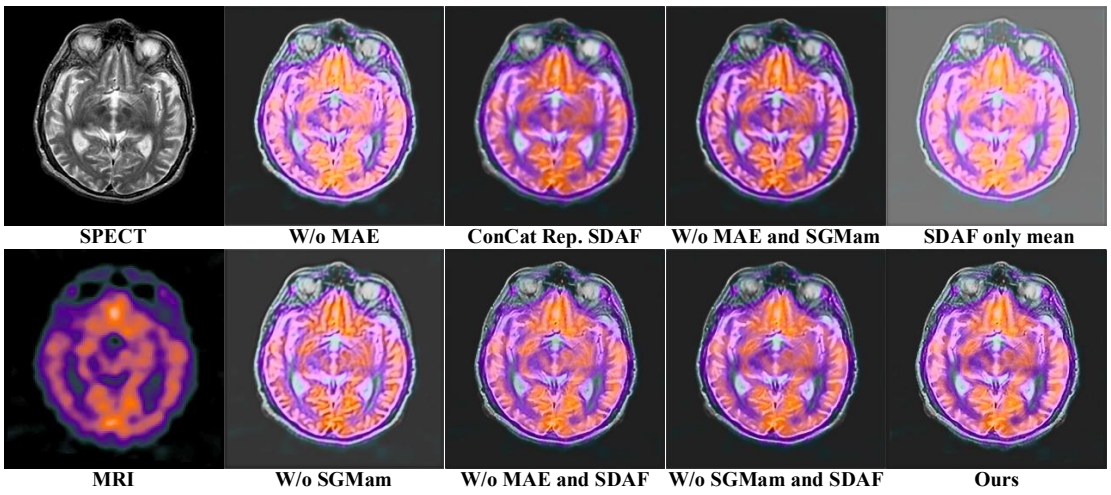


Fig.7. Subjective visualization of ablation studies on the SPECT-MRI dataset

When the MAE module is removed (W/o MAE), the fused images exhibit weaker fine-detail preservation and less clear boundary delineation, indicating that MAE is beneficial for modality-specific enhancement and local complementary feature extraction. Removing the SGMam module (W/o SGMam) leads to reduced global contrast and less prominent functional responses, suggesting that SGMam plays an important role in long-range dependency modeling and salient-region enhancement. When SDAF is replaced by simple concatenation (ConCat Rep. SDAF), the overall structure is still preserved, but the modality transitions become less smooth, and the structural-functional coupling becomes less coherent, which confirms the effectiveness of statistics-guided adaptive fusion.

The combined ablation settings further reveal the collaborative effects among modules. Compared with single-module ablations, removing two components simultaneously generally results in more pronounced degradation in perceptual quality, reflected in flatter contrast, weaker detail hierarchy, and less natural multimodal integration. In addition, the SDAF-only mean variant performs slightly worse

than the full SDAF design, indicating that first-order and second-order statistical descriptors provide complementary guidance for adaptive fusion.

Table 1. Objective results of the network ablation experiments

Datasets	Network	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
CT-MRI	W/o MAE	0.735 [0.684, 0.785]	0.809 [0.799, 0.818]	0.162 [0.151, 0.173]	0.960 [0.953, 0.966]	11.991 [11.765, 12.218]	0.547 [0.507, 0.586]
	W/o SGMam	0.771 [0.737, 0.805]	0.793 [0.783, 0.803]	0.176 [0.160, 0.192]	0.951 [0.946, 0.957]	11.614 [11.399, 11.829]	0.529 [0.492, 0.565]
	Concat Rep. SDAF	0.715 [0.670, 0.760]	0.809 [0.799, 0.819]	0.106 [0.102, 0.111]	0.860 [0.835, 0.885]	12.224 [12.010, 12.439]	0.552 [0.509, 0.594]
	W/o MAE and SDAF	0.714 [0.665, 0.764]	0.805 [0.795, 0.815]	0.104 [0.099, 0.108]	0.832 [0.810, 0.853]	12.454 [12.215, 12.693]	0.537 [0.494, 0.580]
	W/o MAE and SGMam	0.647 [0.605, 0.690]	0.813 [0.802, 0.825]	0.169 [0.161, 0.177]	0.833 [0.792, 0.874]	12.671 [12.439, 12.904]	0.557 [0.515, 0.599]
	W/o SGMam and SDAF	0.678 [0.633, 0.723]	0.807 [0.797, 0.816]	0.151 [0.139, 0.163]	0.793 [0.770, 0.815]	12.603 [12.398, 12.809]	0.562 [0.524, 0.599]
	SDAF only mean	0.458 [0.433, 0.484]	0.733 [0.722, 0.744]	0.112 [0.095, 0.130]	0.986 [0.983, 0.988]	7.602 [7.496, 7.709]	0.493 [0.461, 0.525]
	Ours	0.800 [0.783, 0.817]	0.798 [0.788, 0.808]	0.183 [0.179, 0.187]	0.840 [0.821, 0.859]	13.270 [13.188, 13.352]	0.575 [0.560, 0.590]
PET-MRI	W/o MAE	0.946 [0.914, 0.978]	0.825 [0.811, 0.839]	0.119 [0.112, 0.126]	0.929 [0.914, 0.944]	12.355 [12.096, 12.613]	0.667 [0.655, 0.678]
	W/o SGMam	0.823 [0.792, 0.853]	0.811 [0.796, 0.826]	0.125 [0.117, 0.133]	0.935 [0.922, 0.947]	11.864 [11.529, 12.200]	0.661 [0.650, 0.672]
	Concat Rep. SDAF	0.901 [0.868, 0.934]	0.821 [0.807, 0.834]	0.068 [0.064, 0.072]	0.903 [0.883, 0.924]	12.855 [12.592, 13.118]	0.683 [0.672, 0.695]
	W/o MAE and SDAF	0.947 [0.912, 0.982]	0.828 [0.816, 0.839]	0.072 [0.068, 0.076]	0.877 [0.858, 0.896]	12.614 [12.365, 12.862]	0.669 [0.656, 0.683]
	W/o MAE and SGMam	0.932 [0.897, 0.967]	0.834 [0.821, 0.848]	0.135 [0.127, 0.144]	0.889 [0.863, 0.914]	12.259 [12.016, 12.503]	0.681 [0.668, 0.693]
	W/o SGMam and SDAF	0.907 [0.881, 0.933]	0.831 [0.820, 0.843]	0.112 [0.106, 0.117]	0.881 [0.859, 0.904]	12.425 [12.196, 12.654]	0.682 [0.669, 0.694]
	SDAF only mean	1.012 [0.975, 1.049]	0.833 [0.820, 0.846]	0.146 [0.138, 0.154]	0.885 [0.866, 0.904]	12.559 [12.299, 12.820]	0.676 [0.664, 0.689]
	Ours	0.948 [0.930, 0.966]	0.838 [0.828, 0.848]	0.158 [0.154, 0.162]	0.887 [0.874, 0.900]	13.016 [12.806, 13.226]	0.669 [0.657, 0.681]
SPECT-MRI	W/o MAE	2.083 [1.816, 2.350]	0.877 [0.870, 0.885]	0.147 [0.137, 0.156]	0.919 [0.903, 0.935]	11.758 [11.538, 11.977]	0.676 [0.663, 0.690]
	W/o SGMam	1.469 [1.333, 1.606]	0.848 [0.838, 0.857]	0.122 [0.115, 0.129]	0.924 [0.908, 0.941]	11.088 [10.832, 11.345]	0.657 [0.622, 0.692]
	Concat Rep. SDAF	1.961 [1.748, 2.173]	0.882 [0.874, 0.890]	0.086 [0.080, 0.092]	0.839 [0.823, 0.855]	13.466 [13.215, 13.717]	0.710 [0.695, 0.726]
	W/o MAE and SDAF	1.808 [1.610, 2.006]	0.877 [0.869, 0.885]	0.082 [0.077, 0.088]	0.831 [0.815, 0.846]	14.188 [13.942, 14.434]	0.712 [0.698, 0.727]
	W/o MAE and SGMam	1.250 [1.069, 1.431]	0.883 [0.873, 0.894]	0.136 [0.128, 0.145]	0.672 [0.631, 0.713]	14.971 [14.706, 15.235]	0.706 [0.687, 0.764]
	W/o SGMam and SDAF	1.360 [1.268, 1.451]	0.876 [0.867, 0.884]	0.135 [0.128, 0.143]	0.704 [0.684, 0.724]	14.449 [14.237, 14.662]	0.705 [0.688, 0.722]
	SDAF only mean	0.804 [0.706, 0.902]	0.840 [0.829, 0.851]	0.139 [0.127, 0.150]	0.954 [0.940, 0.969]	7.418 [7.323, 7.513]	0.653 [0.619, 0.688]
	Ours	1.130 [0.959, 1.301]	0.884 [0.873, 0.895]	0.161 [0.156, 0.166]	0.732 [0.714, 0.750]	15.400 [15.020, 15.780]	0.714 [0.698, 0.730]

By contrast, the complete MESFusion model consistently achieves the best visual performance, producing fused images with clearer structural details, more natural modality transitions, better functional expression, cleaner backgrounds, and fewer visible artifacts. These observations demonstrate that the proposed modules are both individually effective and mutually complementary, and that their synergistic integration is essential for high-quality multimodal medical image fusion.

Table 1 reports the quantitative results of the network ablation study on the CT-MRI, PET-MRI, and SPECT-MRI datasets. The full MESFusion model achieves the most competitive and balanced performance across most key metrics. Its strengths are especially clear in QAB/F, PSNR, and MS-SSIM. This demonstrates the effectiveness of the proposed design in preserving structural fidelity, enhancing cross-modal information transfer, and improving perceptual consistency. Some ablated variants may achieve locally higher values on individual indicators such as EN or VIF. However, these gains are not

consistently accompanied by improvements in structural similarity or fusion quality. This suggests that the complete model provides a better overall trade-off between information richness and stable multimodal integration.

From the perspective of single-module ablations, removing MAE generally results in lower PSNR and weaker MS-SSIM scores, indicating that the modality-aware enhancement mechanism contributes to the preservation of fine details and the accurate reconstruction of anatomical structures. Excluding SGMam results in noticeable degradation in global consistency-related metrics, especially in contrast-sensitive and correlation-related performance, confirming its importance for long-range dependency modeling and semantic feature aggregation. Replacing SDAF with a simple concatenation operation results in less stable performance across datasets and weaker improvements in structure- and detail-oriented metrics, further verifying the necessity of the proposed statistics-guided adaptive fusion strategy.

Overall, the quantitative results in Table 1 consistently support the qualitative observations in Figs. 5, 6, and 7, demonstrating that the proposed MESFusion framework achieves superior robustness and a more favorable balance between structural preservation, functional information retention, and cross-modal consistency in multimodal medical image fusion.

4.2.2. Ablation study on loss function hyperparameters

Figs. 8, 9, 10 present the qualitative results of the loss-function hyperparameter ablation experiments on the CT-MRI, PET-MRI, and SPECT-MRI datasets, respectively. The visual comparisons show that different loss terms contribute to complementary aspects of fusion quality, including anatomical structure preservation, contrast enhancement, and functional information consistency.

When the weight of the intensity reconstruction term is reduced or poorly balanced, the fused images tend to exhibit weaker anatomical boundary clarity and less distinct preservation of soft-tissue texture, indicating that this term is important for maintaining spatial structural fidelity. When the gradient-related term is removed, the fused results exhibit reduced local contrast and weaker edge sharpness, resulting in a flatter structural appearance and a less distinct grayscale hierarchy. In contrast, removing the structural similarity term results in more pronounced degradation of structural-functional consistency, especially in PET-MRI and SPECT-MRI fusion, where functional regions become less naturally aligned with anatomical structures and the overall perceptual balance is weakened.

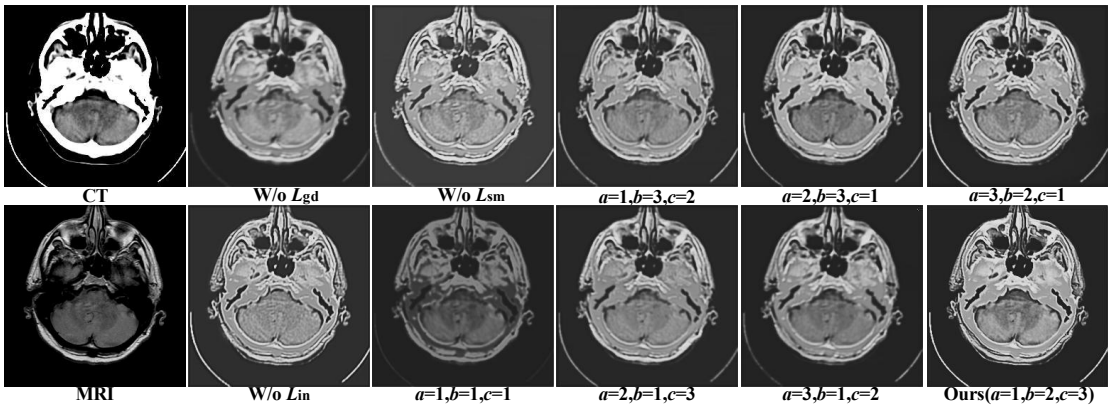


Fig. 8. Subjective results of hyperparameter ablation experiments on the CT-MRI dataset

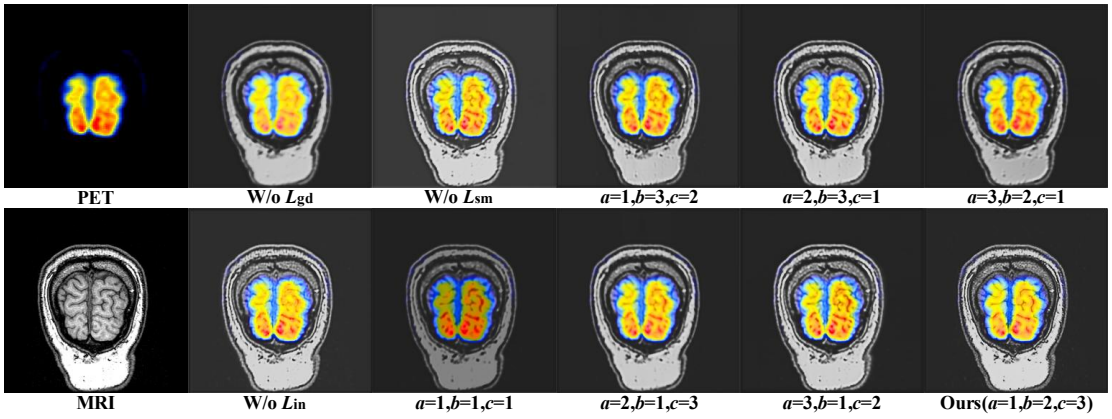


Fig. 9. Subjective results of hyperparameter ablation experiments on the PET-MRI dataset

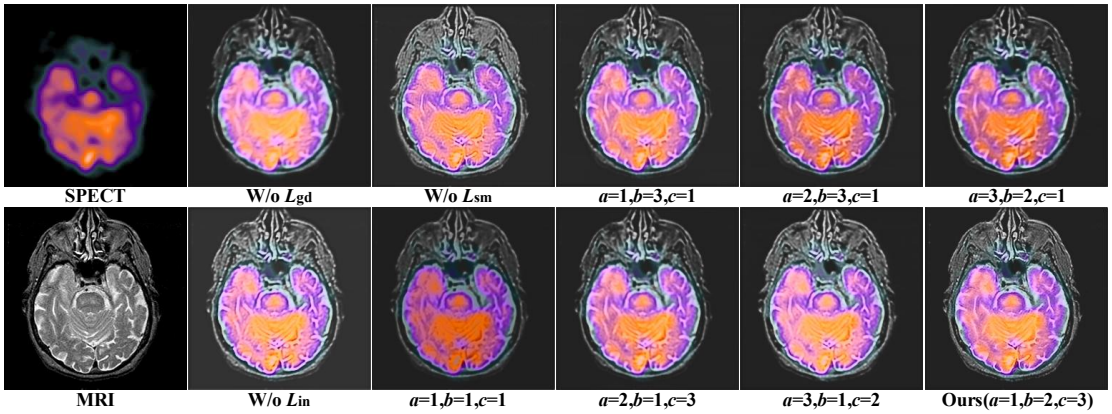


Fig. 10. Subjective results of hyperparameter ablation experiments on the SPECT-MRI dataset

In contrast to other unbalanced or incomplete weighting choices, the proposed hyperparameter configuration a=1, b=2, c=3 consistently yields more visually balanced fusion results across all three datasets. It preserves clearer anatomical details, maintains more natural contrast transitions, and achieves better integration of functional and structural information with fewer visible artifacts. These observations indicate that the selected loss weighting strategy provides a more effective balance among intensity fidelity, structural consistency, and detail preservation.

Tables 2, 3, and 4 report the quantitative results of the loss-function hyperparameter ablation study on the CT-MRI, PET-MRI, and SPECT-MRI datasets, respectively. Overall, the proposed weighting configuration a=1, b=2, c=3 achieves the most competitive, well-balanced performance across key fidelity- and perceptual-quality metrics, particularly in QAB/F, PSNR, and MS-SSIM. Although some alternative settings occasionally yield higher values on individual indicators such as VIF or EN, these improvements are not consistently accompanied by better structural similarity or overall fusion quality. This suggests that the selected weighting strategy provides a more favorable trade-off between information richness, structural preservation, and perceptual consistency.

From the ablation results, removing any individual loss component generally leads to noticeable degradation in the overall metric profile, further verifying that the three loss terms play complementary roles in the optimization process. Specifically, removing the gradient-related term (W/o L_{gd}) tends to weaken edge/detail preservation and perceptual consistency, removing the intensity-consistency term (W/o L_{in}) often affects the maintenance of global grayscale distribution and modality-specific intensity characteristics, and removing the structural similarity term (W/o L_{sm}) usually results in a less stable

balance between structural preservation and information integration. These observations suggest that the three loss components should be optimized collaboratively rather than independently, and that simply emphasizing a single objective cannot reliably achieve robust multimodal fusion.

Table 2 Ablation study of the loss function on the CT-MRI dataset

Hyperparameters	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
W/o L_{gd}	0.774 [0.728, 0.819]	0.815 [0.805, 0.825]	0.177 [0.163, 0.190]	0.929 [0.919, 0.940]	12.885 [12.688, 13.082]	0.557 [0.518, 0.596]
W/o L_{in}	0.612 [0.575, 0.649]	0.798 [0.787, 0.808]	0.147 [0.227, 0.267]	0.945 [0.940, 0.949]	12.413 [12.213, 12.613]	0.504 [0.461, 0.547]
W/o L_{sm}	0.705 [0.672, 0.739]	0.799 [0.789, 0.809]	0.123 [0.119, 0.128]	0.956 [0.949, 0.983]	11.398 [11.197, 11.599]	0.531 [0.492, 0.570]
a=1, b=1, c=1	0.304 [0.290, 0.318]	0.790 [0.779, 0.801]	0.136 [0.118, 0.154]	0.897 [0.378, 0.417]	13.022 [12.718, 13.326]	0.510 [0.473, 0.548]
a=1, b=3, c=2	0.651 [0.618, 0.683]	0.814 [0.805, 0.824]	0.168 [0.157, 0.178]	0.713 [0.700, 0.726]	12.854 [12.669, 13.039]	0.548 [0.507, 0.589]
a=2, b=1, c=3	0.662 [0.626, 0.697]	0.816 [0.806, 0.825]	0.178 [0.163, 0.193]	0.879 [0.862, 0.897]	13.174 [12.968, 13.379]	0.545 [0.504, 0.585]
a=2, b=3, c=1	0.680 [0.636, 0.723]	0.806 [0.796, 0.816]	0.157 [0.150, 0.164]	0.795 [0.772, 0.818]	13.179 [12.978, 13.381]	0.551 [0.509, 0.592]
a=3, b=1, c=2	0.768 [0.726, 0.809]	0.805 [0.795, 0.814]	0.144 [0.138, 0.151]	0.886 [0.873, 0.899]	12.925 [12.712, 13.138]	0.537 [0.494, 0.579]
a=3, b=2, c=1	0.625 [0.587, 0.662]	0.808 [0.799, 0.818]	0.177 [0.192, 0.221]	0.763 [0.741, 0.784]	13.779 [13.570, 13.987]	0.557 [0.519, 0.595]
Ours (a=1, b=2, c=3)	0.800 [0.783, 0.817]	0.798 [0.788, 0.808]	0.183 [0.179, 0.187]	0.840 [0.821, 0.859]	13.270 [13.188, 13.352]	0.575 [0.560, 0.590]

Table 3 Ablation study of the loss function on the SPECT-MRI dataset

Hyperparameters	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
W/o L_{gd}	0.942 [0.911, 0.973]	0.823 [0.811, 0.875]	0.117 [0.111, 0.124]	0.881 [0.862, 0.901]	12.929 [12.675, 13.184]	0.665 [0.653, 0.677]
W/o L_{in}	0.833 [0.795, 0.871]	0.809 [0.796, 0.821]	0.219 [0.207, 0.232]	0.825 [0.810, 0.840]	12.363 [12.114, 12.611]	0.650 [0.637, 0.664]
W/o L_{sm}	1.004 [0.972, 1.036]	0.816 [0.802, 0.831]	0.088 [0.084, 0.091]	0.835 [0.819, 0.850]	11.785 [11.552, 12.019]	0.666 [0.634, 0.697]
a=1, b=1, c=1	0.369 [0.344, 0.393]	0.825 [0.812, 0.838]	0.208 [0.198, 0.218]	0.629 [0.612, 0.646]	12.756 [12.572, 12.939]	0.597 [0.584, 0.609]
a=1, b=3, c=2	1.003 [0.969, 1.038]	0.827 [0.815, 0.840]	0.124 [0.117, 0.131]	0.857 [0.836, 0.878]	13.001 [12.761, 13.242]	0.656 [0.634, 0.678]
a=2, b=1, c=3	0.940 [0.901, 0.979]	0.830 [0.817, 0.843]	0.131 [0.123, 0.140]	0.881 [0.861, 0.901]	12.572 [12.311, 12.834]	0.668 [0.645, 0.670]
a=2, b=3, c=1	1.023 [0.985, 1.060]	0.816 [0.802, 0.829]	0.119 [0.112, 0.125]	0.882 [0.863, 0.900]	12.175 [11.902, 12.448]	0.654 [0.641, 0.667]
a=3, b=1, c=2	1.064 [1.031, 1.096]	0.820 [0.806, 0.834]	0.110 [0.105, 0.115]	0.807 [0.789, 0.825]	12.852 [12.601, 13.102]	0.664 [0.652, 0.677]
a=3, b=2, c=1	0.971 [0.939, 1.002]	0.826 [0.813, 0.839]	0.153 [0.144, 0.162]	0.871 [0.848, 0.894]	12.562 [12.309, 12.816]	0.655 [0.642, 0.668]
Ours (a=1, b=2, c=3)	0.948 [0.930, 0.966]	0.838 [0.828, 0.848]	0.158 [0.154, 0.162]	0.887 [0.874, 0.900]	13.016 [12.806, 13.226]	0.669 [0.657, 0.681]

Table 4 Ablation study of the loss function on the PET-MRI dataset

Hyperparameters	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
W/o L_{gd}	1.803 [1.603, 2.003]	0.894 [0.887, 0.900]	0.128 [0.121, 0.135]	0.887 [0.870, 0.905]	13.134 [12.925, 13.343]	0.699 [0.685, 0.714]
W/o L_{in}	1.720 [1.525, 1.915]	0.870 [0.862, 0.878]	0.207 [0.192, 0.221]	0.911 [0.867, 0.956]	12.234 [12.017, 12.452]	0.670 [0.655, 0.685]
W/o L_{sm}	1.872 [1.644, 2.101]	0.868 [0.850, 0.885]	0.118 [0.112, 0.124]	0.914 [0.897, 0.931]	11.347 [11.166, 11.527]	0.682 [0.668, 0.697]
a=1, b=1, c=1	0.741 [0.659, 0.823]	0.879 [0.870, 0.887]	0.191 [0.181, 0.200]	0.456 [0.440, 0.473]	14.943 [14.718, 15.168]	0.691 [0.675, 0.708]
a=1, b=3, c=2	1.630 [1.460, 1.800]	0.889 [0.871, 0.906]	0.129 [0.120, 0.139]	0.867 [0.848, 0.885]	13.258 [13.051, 13.464]	0.711 [0.695, 0.728]
a=2, b=1, c=3	1.672 [1.500, 1.844]	0.861 [0.844, 0.878]	0.122 [0.116, 0.128]	0.857 [0.841, 0.874]	13.695 [13.474, 13.915]	0.714 [0.698, 0.729]
a=2, b=3, c=1	1.418 [1.270, 1.567]	0.886 [0.868, 0.903]	0.117 [0.111, 0.123]	0.782 [0.746, 0.818]	14.505 [14.259, 14.750]	0.705 [0.689, 0.722]
a=3, b=1, c=2	2.046 [1.801, 2.291]	0.882 [0.874, 0.890]	0.136 [0.128, 0.144]	0.882 [0.838, 0.927]	12.626 [12.417, 12.835]	0.685 [0.640, 0.730]
a=3, b=2, c=1	1.422 [1.286, 1.557]	0.882 [0.874, 0.890]	0.139 [0.131, 0.147]	0.746 [0.729, 0.764]	15.181 [14.955, 15.407]	0.706 [0.689, 0.712]
Ours (a=1, b=2, c=3)	1.130 [0.959, 1.301]	0.884 [0.873, 0.895]	0.161 [0.156, 0.166]	0.732 [0.714, 0.750]	15.400 [15.020, 15.780]	0.714 [0.698, 0.730]

4.3. Comparative experiments

To comprehensively and objectively evaluate the performance of our fusion network, we selected seven advanced methods from recent years for comparison: U2Fusion (22, TPAMI) [24], SwinFusion (22, ECCV) [19], CDDFuse (23, CVPR) [25], FusionMamba (24, VI) [11], SAFFusion (25, BOE) [26], SMAFusion (25, NCOM) [27], MARCFusion (25, MS) [28], DDBFusion (26, IF)[29] and DM-FNet (TMM 26)[30].

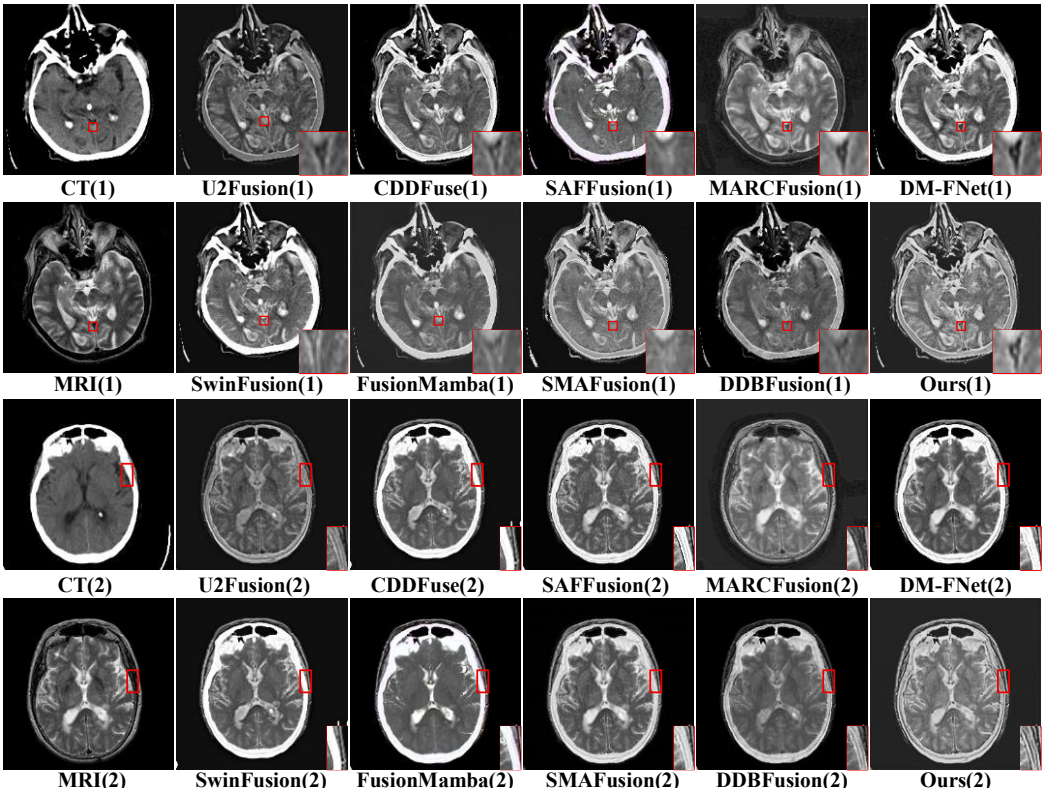


Fig. 11. Subjective results of comparative experiments on the CT-MRI dataset

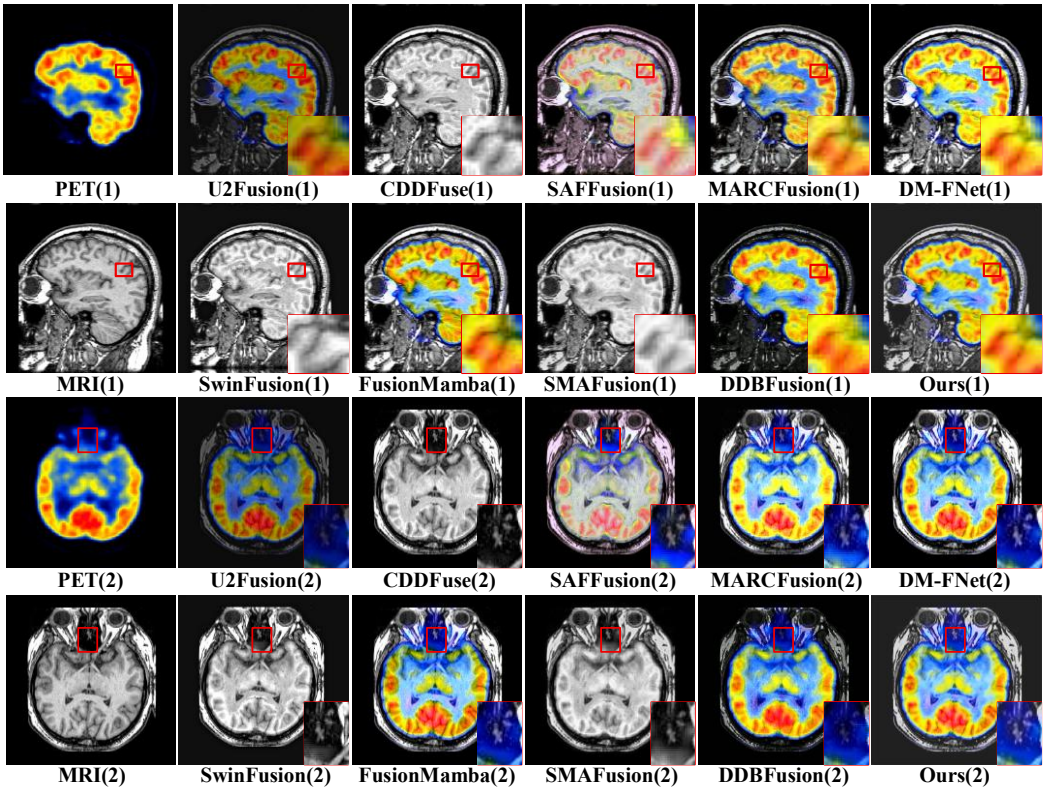


Fig. 12. Subjective results of comparative experiments on the PET-MRI dataset

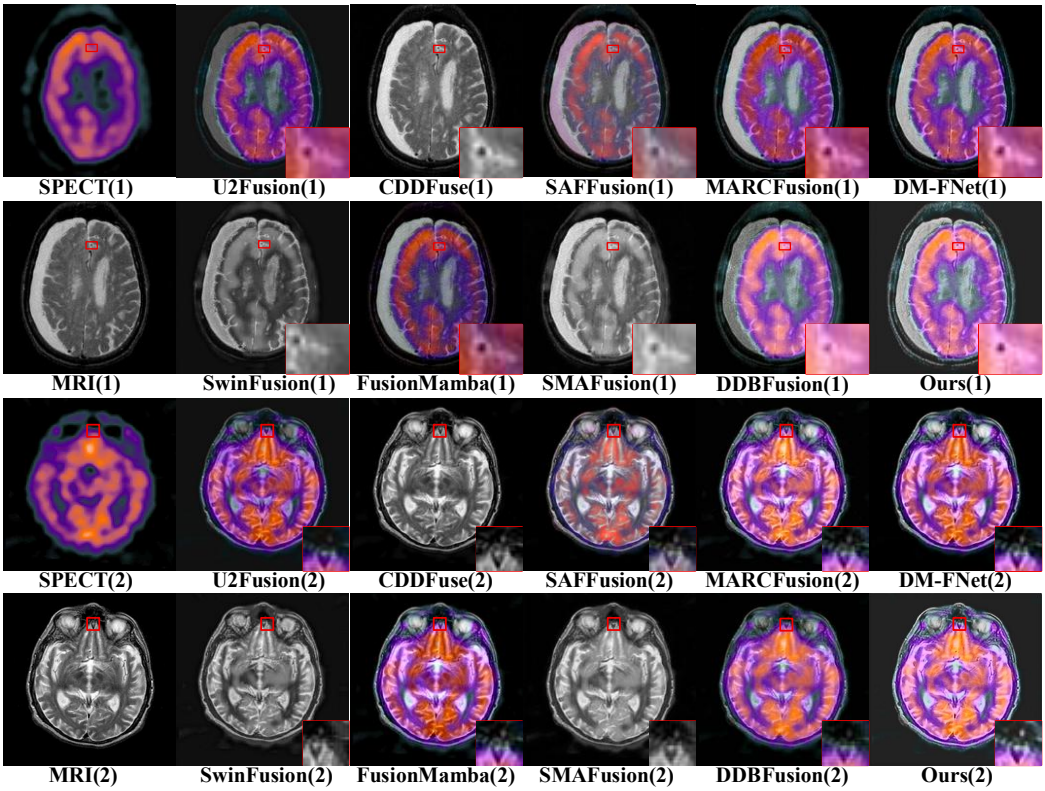


Fig. 13. Subjective results of comparative experiments on the SPECT-MRI dataset

Figs. 11, 12, and 13 present qualitative comparisons between the proposed method and representative baseline methods on the CT-MRI, PET-MRI, and SPECT-MRI datasets, respectively. Although existing fusion methods can integrate cross-modal information to some extent, achieving a balanced preservation of both structural and functional information remains challenging. In particular, some methods exhibit blurred anatomical boundaries due to information loss during feature extraction, while others suffer from reduced contrast or a weakened representation of salient functional regions during fusion. In comparison, the proposed method produces visually more balanced fusion results. Benefiting from the collaborative design of the MAE, SGMam, and SDAF modules, MESFusion effectively extracts and integrates complementary information from heterogeneous modalities while preserving structural continuity and reducing fusion artifacts. The resulting fused images exhibit clearer anatomical details, more natural intensity transitions, and more consistent representation of salient functional information. Overall, the visual comparisons indicate that the proposed method achieves a favorable visual quality and a better balance between structural preservation and information integration across different multimodal medical image fusion tasks.

To further examine the generalization ability of the proposed method, Fig. 14 presents subjective comparison results on the NAF-PROSTATE dataset, an external PET/CT dataset with a different distribution from the training data. It can be observed that several comparative methods exhibit noticeable performance degradation when directly transferred to this unseen dataset, including local structural blurring, reduced lesion contrast, and less stable preservation of fine anatomical details in enlarged regions. In comparison, the proposed method still produces fused images with clearer structural boundaries, more stable background textures, and better preservation of lesion-related high-intensity regions. This suggests that the proposed framework maintains relatively robust fusion quality across datasets and exhibits favorable generalization performance in more challenging cross-dataset scenarios.

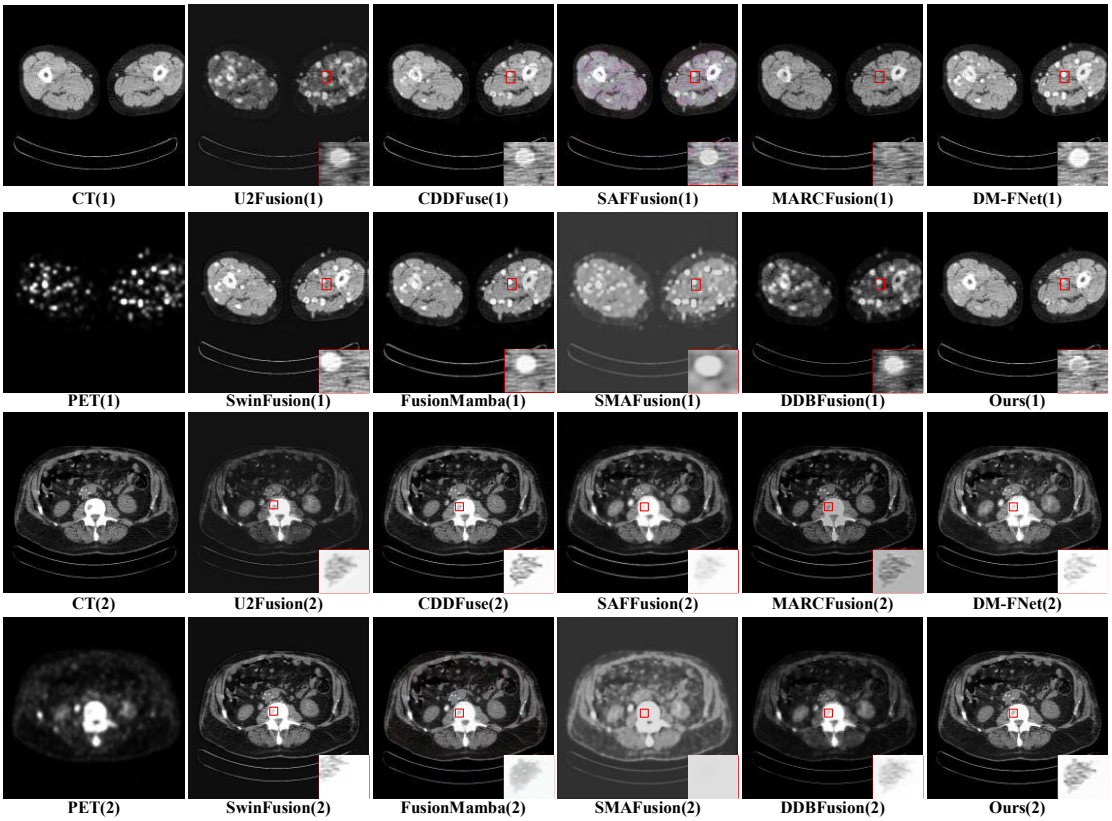


Fig. 14. Subjective comparison results on the external NAF-PROSTATE dataset

Tables 5, 6, and 7 report the quantitative comparison results of different methods on the CT-MRI, PET-MRI, and SPECT-MRI datasets, respectively. Overall, the proposed method demonstrates strong and consistent performance across the three multimodal medical image fusion tasks, achieving highly competitive results on several representative structure- and fidelity-related metrics, including CC, PSNR, and MS-SSIM. Although some competing methods obtain higher values on individual indicators such as EN or VIF under specific settings, these improvements are often not accompanied by comparable structural fidelity or perceptual consistency. Consequently, the proposed method achieves a more favorable balance among information preservation, structural continuity, and overall visual quality. The effectiveness of MESFusion can be primarily attributed to the collaborative design of its key components. First, the Modality-Aware Enhancement (MAE) module performs modality-specific feature refinement through multi-scale gradient extraction and intensity statistical modeling, which helps preserve anatomical details while enhancing complementary information from heterogeneous modalities. Second, the SGMam module introduces selective state-space modeling to capture long-range spatial dependencies and global contextual interactions with linear computational complexity, thereby improving structural consistency and feature representation capability. Furthermore, the Statistical-Guided Adaptive Fusion (SDAF) module dynamically adjusts the fusion process according to the statistical properties of heterogeneous features, enabling a more effective balance between structural texture and salient information while reducing artifacts and information suppression. Finally, the composite loss function jointly constrains intensity fidelity, gradient preservation, and structural similarity, further promoting the generation of visually coherent fusion results with clear structural details and enhanced feature representation.

Table 5 Objective results of comparative experiments on the CT-MRI

Methods	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
U2Fusion	0.585 [0.548, 0.622]	0.788 [0.777, 0.800]	0.114 [0.107, 0.121]	0.566 [0.524, 0.608]	15.344 [15.036, 15.651]	0.567 [0.548, 0.587]

SwinFusion	0.574 [0.533, 0.616]	0.639 [0.618, 0.660]	0.177 [0.168, 0.206]	0.739 [0.700, 0.779]	11.631 [11.339, 11.924]	0.407 [0.389, 0.426]
CDDFuse	1.337 [1.264, 1.410]	0.795 [0.781, 0.808]	0.082 [0.077, 0.088]	0.746 [0.706, 0.786]	13.923 [13.574, 14.272]	0.545 [0.501, 0.590]
FusionMamba	1.375 [1.305, 1.446]	0.777 [0.765, 0.788]	0.122 [0.112, 0.132]	0.695 [0.647, 0.743]	13.452 [13.095, 13.809]	0.534 [0.469, 0.569]
SMAFusion	0.758 [0.712, 0.804]	0.792 [0.780, 0.804]	0.178 [0.169, 0.187]	0.821 [0.779, 0.863]	13.738 [13.454, 14.022]	0.519 [0.491, 0.576]
SAFFusion	0.807 [0.793, 0.820]	0.737 [0.724, 0.749]	0.137 [0.120, 0.154]	0.729 [0.684, 0.773]	12.738 [12.515, 12.961]	0.492 [0.475, 0.508]
MARCFusion	0.751 [0.717, 0.786]	0.773 [0.760, 0.787]	0.114 [0.108, 0.121]	0.717 [0.668, 0.766]	13.380 [13.078, 13.683]	0.564 [0.541, 0.588]
DDBFusion	0.760 [0.718, 0.801]	0.731 [0.719, 0.744]	0.122 [0.112, 0.133]	0.621 [0.579, 0.664]	15.534 [15.180, 15.887]	0.574 [0.571, 0.617]
DM-FNet	1.519 [1.434, 1.604]	0.789 [0.777, 0.801]	0.094 [0.089, 0.099]	0.781 [0.744, 0.818]	13.259 [12.919, 13.600]	0.545 [0.502, 0.587]
Ours	0.800 [0.783, 0.817]	0.798 [0.788, 0.808]	0.183 [0.179, 0.187]	0.840 [0.821, 0.859]	13.270 [13.188, 13.352]	0.575 [0.560, 0.590]

Table 6 Objective results of comparative experiments on the PET-MRI

Methods	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
U2Fusion	0.450 [0.431, 0.469]	0.821 [0.811, 0.831]	0.101 [0.096, 0.106]	0.307 [0.271, 0.343]	14.234 [13.936, 14.459]	0.624 [0.617, 0.646]
SwinFusion	0.563 [0.521, 0.604]	0.628 [0.607, 0.649]	0.149 [0.137, 0.181]	0.863 [0.839, 0.887]	10.698 [10.469, 11.043]	0.418 [0.391, 0.434]
CDDFuse	1.588 [1.570, 1.606]	0.787 [0.771, 0.803]	0.044 [0.041, 0.048]	0.879 [0.856, 0.902]	12.579 [12.357, 13.033]	0.658 [0.644, 0.697]
FusionMamba	1.145 [1.104, 1.185]	0.791 [0.778, 0.805]	0.079 [0.072, 0.086]	0.805 [0.780, 0.830]	13.887 [13.560, 14.214]	0.650 [0.649, 0.675]
SMAFusion	1.363 [1.324, 1.403]	0.806 [0.797, 0.829]	0.039 [0.036, 0.043]	0.950 [0.927, 0.974]	9.311 [9.121, 9.501]	0.628 [0.612, 0.690]
SAFFusion	0.612 [0.584, 0.639]	0.757 [0.744, 0.770]	0.105 [0.098, 0.112]	0.863 [0.849, 0.877]	12.794 [12.669, 13.454]	0.663 [0.650, 0.677]
MARCFusion	1.354 [1.311, 1.397]	0.795 [0.763, 0.797]	0.044 [0.040, 0.048]	0.864 [0.840, 0.888]	13.295 [12.831, 13.760]	0.660 [0.656, 0.681]
DDBFusion	1.047 [1.008, 1.086]	0.812 [0.801, 0.823]	0.082 [0.077, 0.086]	0.815 [0.784, 0.846]	15.172 [14.847, 15.497]	0.658 [0.646, 0.700]
DM-FNet	1.676 [1.622, 1.729]	0.809 [0.793, 0.824]	0.040 [0.037, 0.044]	0.878 [0.855, 0.902]	12.695 [12.360, 13.030]	0.663 [0.661, 0.695]
Ours	0.948 [0.930, 0.966]	0.838 [0.828, 0.848]	0.158 [0.154, 0.162]	0.887 [0.874, 0.900]	13.016 [12.806, 13.226]	0.669 [0.657, 0.681]

Table 7 Objective results of comparative experiments on the SPECT-MRI

Methods	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
U2Fusion	0.740 [0.665, 0.814]	0.866 [0.848, 0.885]	0.090 [0.084, 0.096]	0.124 [0.090, 0.157]	19.789 [19.212, 20.366]	0.686 [0.663, 0.710]
SwinFusion	0.895 [0.742, 1.048]	0.739 [0.698, 0.779]	0.177 [0.136, 0.197]	0.453 [0.366, 0.541]	16.730 [16.238, 17.221]	0.496 [0.475, 0.517]
CDDFuse	2.206 [1.887, 2.526]	0.801 [0.765, 0.837]	0.058 [0.049, 0.067]	0.493 [0.397, 0.588]	17.120 [16.005, 18.235]	0.580 [0.511, 0.649]
FusionMamba	1.480 [1.282, 1.679]	0.851 [0.827, 0.876]	0.068 [0.057, 0.080]	0.443 [0.338, 0.548]	18.835 [18.100, 19.570]	0.657 [0.603, 0.712]
SMAFusion	1.248 [1.131, 1.364]	0.822 [0.787, 0.857]	0.066 [0.057, 0.076]	0.262 [0.176, 0.347]	19.583 [18.745, 20.420]	0.651 [0.593, 0.709]
SAFFusion	1.771 [1.575, 1.968]	0.820 [0.792, 0.847]	0.055 [0.046, 0.064]	0.425 [0.332, 0.519]	18.318 [17.540, 19.096]	0.604 [0.534, 0.674]
MARCFusion	2.199 [1.867, 2.530]	0.867 [0.841, 0.893]	0.086 [0.076, 0.095]	0.692 [0.592, 0.792]	16.233 [15.087, 17.379]	0.682 [0.653, 0.712]
DDBFusion	1.465 [1.291, 1.640]	0.870 [0.852, 0.889]	0.089 [0.079, 0.098]	0.551 [0.413, 0.688]	18.996 [17.574, 20.418]	0.711 [0.685, 0.745]
DM-FNet	1.911 [1.798, 2.024]	0.869 [0.855, 0.882]	0.054 [0.050, 0.059]	0.600 [0.560, 0.641]	17.731 [17.121, 18.340]	0.696 [0.668, 0.724]
Ours	1.130 [0.959, 1.301]	0.884 [0.873, 0.895]	0.161 [0.156, 0.166]	0.732 [0.714, 0.750]	15.400 [15.020, 15.780]	0.714 [0.698, 0.730]

To further validate the cross-dataset generalization capability of the proposed method, Table 8 reports the quantitative comparison results on the external NAF-PROSTATE PET/CT dataset. Despite a shift in distribution between the training datasets and the unseen external dataset, the proposed method still achieves the best performance on several key structure- and fidelity-related metrics, including CC, PSNR, and MS-SSIM, demonstrating strong robustness under cross-dataset transfer. Although certain comparative methods obtain higher values on individual indicators such as VIF, EN, or QAB/F, these improvements do not necessarily correspond to better structural consistency or more stable visual quality, as also reflected in the subjective results shown in Fig. 14. In contrast, the proposed method achieves a more balanced overall performance, indicating that it can more reliably preserve anatomical structures while maintaining salient functional information in unseen datasets. These results further confirm the cross-dataset generalization capability of the proposed framework beyond the in-domain test sets. These results suggest that the proposed method maintains favorable fusion quality and robustness under cross-dataset distribution shifts, indicating promising generalization ability in more cross-dataset imaging scenarios.

Table 9 presents the comparison of Flops and parameter counts for different fusion methods. To ensure fairness, all methods were evaluated with the same input size ($1\times1\times224\times224$) when calculating computational complexity. As shown, the proposed method is well-suited for deployment in resource-constrained environments, owing to its task-oriented, lightweight architecture, efficient global dependency modeling, and compact statistical-guided adaptive fusion design. These characteristics make it particularly suitable for resource-constrained deployment scenarios due to its lightweight design.

Table 8 Objective results of the comparative experiments on the NAF-PROSTATE dataset

Methods	VIF	CC	QAB/F	EN	PSNR	MS-SSIM
U2Fusion	1.340 [1.317, 1.363]	0.931 [0.928, 0.933]	0.071 [0.069, 0.072]	0.294 [0.280, 0.308]	20.555 [20.434, 20.676]	0.853 [0.849, 0.858]
SwinFusion	1.085 [1.051, 1.119]	0.831 [0.826, 0.835]	0.341 [0.337, 0.345]	0.445 [0.432, 0.458]	17.408 [17.274, 17.543]	0.695 [0.688, 0.701]

CDDFuse	1.880 [1.841, 1.919]	0.909 [0.906, 0.913]	0.070 [0.067, 0.072]	0.442 [0.427, 0.456]	20.346 [20.144, 20.547]	0.829 [0.824, 0.834]
FusionMamba	1.866 [1.821, 1.911]	0.918 [0.915, 0.921]	0.079 [0.076, 0.082]	0.432 [0.418, 0.447]	20.513 [20.289, 20.737]	0.837 [0.832, 0.842]
SMAFusion	0.864 [0.831, 0.897]	0.886 [0.883, 0.890]	0.385 [0.371, 0.399]	0.679 [0.670, 0.689]	13.077 [12.982, 13.171]	0.749 [0.745, 0.752]
SAFFusion	1.363 [1.328, 1.399]	0.898 [0.894, 0.901]	0.077 [0.074, 0.080]	0.280 [0.268, 0.291]	20.485 [20.319, 20.651]	0.825 [0.821, 0.830]
MARCFusion	1.849 [1.805, 1.893]	0.903 [0.899, 0.908]	0.076 [0.073, 0.079]	0.448 [0.434, 0.462]	20.066 [19.850, 20.282]	0.827 [0.821, 0.832]
DDBFusion	1.438 [1.413, 1.463]	0.942 [0.939, 0.945]	0.086 [0.083, 0.089]	0.271 [0.260, 0.282]	20.594 [20.388, 20.799]	0.821 [0.817, 0.825]
DM-FNet	2.128 [2.079, 2.178]	0.906 [0.902, 0.909]	0.054 [0.053, 0.056]	0.485 [0.472, 0.498]	19.133 [18.970, 19.297]	0.814 [0.809, 0.819]
Ours	1.835 [1.413, 2.257]	0.945 [0.943, 0.947]	0.066 [0.061, 0.071]	0.451 [0.437, 0.465]	20.836 [20.748, 20.924]	0.856 [0.853, 0.859]

Table 9 Comparison of flops, parameter counts, and inference time across fusion methods

Methods	Flops	Params	Runtime (Average value)
U2Fusion	86.447G	659.217K	1.125
SwinFusion	63.733G	927.452K	0.932
CDDFuse	53.329G	807.018K	0.574
FusionMamba	42.059G	729.303K	0.752
SMAFusion	85.477G	866.000K	1.056
SAFFusion	18.242G	277.821K	0.129
MARCFusion	19.153G	391.746K	0.276
DDBFusion	369.862G	3647.000K	1.596
DM-FNet	298.250G	20.648M	1.232
Ours	3.511G	31.6230K	0.056

5. Conclusion

Addressing the challenges of significant feature distribution discrepancies and insufficient long-range dependency modeling in multimodal medical image fusion, this paper proposes a problem-driven framework that couples modality-specific enhancement with statistically guided adaptive fusion. The proposed Modality-Aware Enhancement (MAE) module employs multi-scale gradient operators and energy statistics to enhance anatomical structures and functional information from different modalities, thereby improving the representation capability of heterogeneous features. The SGMam module further captures long-range spatial dependencies and global contextual information through a selective scan mechanism while maintaining linear computational complexity. In addition, the Statistical-Guided Adaptive Fusion (SDAF) module dynamically calibrates fusion behavior using first- and second-order statistical priors of feature channels, enabling effective integration of complementary cross-modal information. To evaluate the effectiveness of the proposed framework, extensive comparative experiments and ablation studies were conducted on three publicly available multimodal medical image fusion datasets, including CT-MRI, PET-MRI, and SPECT-MRI. Experimental results demonstrate that MESFusion achieves competitive and well-balanced performance across multiple quantitative metrics while maintaining favorable visual quality and structural consistency. Furthermore, cross-dataset evaluation on the external NAF-PROSTATE dataset indicates that the proposed method exhibits promising robustness under distribution shifts. Despite these encouraging results, one limitation of the current framework is that the weighting coefficients of the composite loss function are still manually determined, which increases the empirical tuning burden and experimental cost. Future work will focus on adaptive loss-weight optimization strategies, broader cross-dataset validation, and further improvements in the efficiency and generalization capability of multimodal medical image fusion models.

References

[1] Khan, M. Z., Diwakar, M., Srivastava, P. et al. Multimodality medical image fusion using directional total variation based linear spectral clustering in NSCT domain. *Sci. Rep*. 16, 5367 (2026).

[2] Lin, C., Chen, Y., Feng, S. et al. A multibranch and multiscale neural network based on semantic

perception for multimodal medical image fusion. *Sci. Rep*. 14, 17609 (2024).

[3] Wu, X. et al. CAFNet: Circular Attention Fusion for medical image segmentation. *Knowl.-Based Syst*. 329, 114277 (2025).

[4] Lyu, Y. et al. DDPM-EMF: a denoising diffusion probabilistic model-based feature-enhancement fusion network for medical image fusion. *J. Opt. Soc. Am. A* 42, 549–562 (2025).

[5] Azam, M. A. et al. A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Comput. Biol. Med*. 144, 105253 (2022).

[6] Wei, L. et al. MMIFprompt: Efficient modality-aware prompting for multimodal medical image fusion. *Pattern Recognit*. 171, 112266 (2026).

[7] Zhou, L. et al. Adversarially robust Fourier-aware multimodal medical image fusion for LSCI. *Neurocomputing* 660, 131889 (2026).

[8] Zhou, T. et al. Long-range correlation-guided dual-encoder fusion network for medical images. *Sci. Rep*. 15, 38964 (2025).

[9] Wang, C. et al. DiCyc: GAN-based deformation invariant cross-domain information fusion for medical image synthesis. *Inf. Fusion* 67, 147–160 (2021).

[10] Qin, J. et al. CNN–Transformer gated fusion network for medical image super-resolution. *Sci. Rep*. 15, 15338 (2025).

[11] Xie, X. et al. Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba. *Vis. Intell*. 2, 37 (2024).

[12] Wu, X. et al. W-EICMFusion: A fusion network for infrared and visible images utilising WOA hyperparameter optimisation. Pattern Recognit. 172, 112531 (2026).

[13] Dong, et al. A fusion network for multi-modality medical image registration with progressive feature alignment. *Knowl.-Based Syst*. 317, 113427 (2025).

[14] Zhao, Z. et al. DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion. In Proc. IEEE/CVF Int. *Conf. Comput. Vis.* 8048–8059 (2023).

[15] Guan, X. et al. EMFF-Net: Edge-Enhancement Multi-Scale Feature Fusion Network. *IEEE Access* 13, 25598–25611 (2025).

[16] Chao, X. et al. Constrained Generative Adversarial Networks. *IEEE Access* 9, 19208–19218 (2021).

[17] Li, H. et al. Hybrid generative adversarial network based on a mixed attention fusion module for multi-modal MR image synthesis algorithm. Int. J. *Mach. Learn*. Cybern. 15, 2111–2130 (2024).

[18] Tang, W. et al. MATR: Multimodal Medical Image Fusion via Multiscale Adaptive Transformer. *IEEE Trans. Image Process.* 31, 5134–5149 (2022).

[19] Ma, J. et al. SwinFusion: Cross-domain Long-range Learning for General Image Fusion via Swin Transformer. *IEEE/CAA J. Autom. Sinica* 9, 1200–1217 (2022).

[20] Li, Z. et al. MambaDFuse: A Mamba-based Dual-phase Model for Multi-modality Image Fusion. *arXiv preprint arXiv:*2404.08406 (2024).

[21] Wu, X. & Feng, X. DTEFusion: a dual-transformation network enhanced fusion method for infrared and visible images with improved mamba. *Infrared Phys. Technol*. 157, 106647 (2026).

[22] Kurdziel, K. A. et al. Data From NaF PROSTATE [Dataset]. The Cancer Imaging Archive (2015).

[23] Wu, X. et al. A Method for Infrared and Visible Image Fusion Based on CNN-Mamba Feature Extraction Combined with Information Selection. *Acta Photonica Sinica* 55, 0110001 (2026).

[24] Xu, H. et al. U2Fusion: A Unified Unsupervised Image Fusion Network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 5021–5037 (2022).

[25] Zhao, Z. et al. CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion. In Proc. *IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 5906–5916 (2023).
[26] Liu, R. et al. SAFFusion: a saliency-aware frequency fusion network for multimodal medical image fusion. *Biomed. Opt. Express* 16, 2459–2481 (2025).
[27] Guo, W. et al. SMAFusion: Multimodal medical image fusion based on spatial registration and local-global multi-scale feature adaptive fusion. *Neurocomputing* 652, 131039 (2025).
[28] Tang, H. et al. MARCFusion: adaptive residual cross-domain fusion network for medical image fusion. *Multimed. Syst*. 31, 99 (2025).
[29] Zhang, Z. et al. DDBFusion: An unified image decomposition and fusion framework based on dual decomposition and Bézier curves. *Inf. Fusion* 114, 102655 (2025).
[30] He, D. et al. DM-FNet: Unified multimodal medical image fusion via diffusion process-trained encoder-decoder. *IEEE Trans. Multimedia* 27, 9415–9428 (2025).

Funding

This work was supported by the Jiamusi University New Medical Science Medical Education Special Project (Grant No. 2026XYK-13).

Datasets

Datasets
The datasets used in this study are publicly available. The SPECT-MRI, CT-MRI, and PET-MRI datasets can be accessed from the Harvard Medical Image Fusion Dataset repository: https://github.com/xianming-gu/Havard-Medical-Image-Fusion-Datasets.

In addition, the NAF-Prostate dataset is available from The Cancer Imaging Archive (TCIA): https://www.cancerimagingarchive.net/collection/naf-prostate/.
The NAF-Prostate dataset (TCIA) is available at:
https://www.cancerimagingarchive.net/collection/naf-prostate/