Neural Networks 202 (2026) 108994

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Full Length Article

FDPMambaFuse: A frequency-domain and parallel Mamba-based model for multimodal medical image fusion

Baiya Li [a,1], Boheng Zhang [b,c,d,e,1], Yang Liu [b], Haorui Huang [b,d], Cailing Lin [f], Rui Fan [g], Mingjian Sun [b,d,f,h,e,*]

[a] Department of Otolaryngology Head and Neck Surgery, The First Affiliated Hospital of Xi'an Jiaotong University, Xi'an, 710061, China
[b] Department of Control Science and Engineering, Harbin Institute of Technology, Harbin, 150001, China
[c] Department of Convergence IT Engineering, Pohang University of Sicence and Technology, Pohang, 37673, Republic of Korea
[d] Harbin Institute of Technology Suzhou Research Institute, Suzhou, 215000, China
[e] Harbin Institute of Technology (Weihai) Qingdao Research Institute, Qingdao, 266109, China
[f] Department of Information Science and Engineering, Harbin Institute of Technology, Weihai, 264200, China
[g] Machine Intelligence & Autonomous Systems (MIAS) Group, the College of Electronics & Information Engineering, Tongji University, Shanghai, 201804, China
[h] Shandong Laboratory of Advanced Biomaterials and Medical Devices in Weihai, Weihai, 264209, China

ARTICLE INFO

Keywords:
Multimodal medical image fusion
Frequency-domain feature modeling
Multi-directional vision Mamba
Unsupervised learning
Photoacoustic-ultrasound fusion

ABSTRACT

Multimodal medical image fusion (MMIF) integrates complementary information from different imaging modalities, providing more comprehensive and reliable support for clinical diagnosis and analysis. Although Mamba-based models efficiently capture long-range dependencies with low computational cost, existing approaches often overlook frequency-domain features and directional structural information, resulting in insufficient detail preservation and reduced semantic consistency. To address these limitations, we propose FDPMambaFuse, a fusion network that combines frequency-domain decomposition with multi-directional parallel Mamba. The framework employs a multi-level CNN-Mamba feature extractor, in which CNNs capture shallow spatial features, whereas a discrete wavelet transform (DWT)-based parallel Mamba module models multi-scale frequency components, long-range dependencies, and direction-sensitive structural details. A dual-stage fusion module is subsequently introduced to enhance cross-modal complementarity through channel interaction and spatially adaptive attention. In addition, we design an unsupervised hybrid loss that combines pixel-level uncertainty with image gradients to improve structural consistency and overall visual quality. To further improve model generalization, we construct and publicly release a high-quality photoacoustic-ultrasound fusion dataset, HIT-MMIF-PAUS. Extensive qualitative and quantitative experiments demonstrate that FDPMambaFuse outperforms state-of-the-art methods in fusion quality, structural fidelity, and edge clarity. Moreover, it achieves superior accuracy and robustness in downstream tumor segmentation tasks, further verifying its practical potential in clinical applications. Our code is publicly available at https://github.com/670768312/FDPMambaFuse.

1. Introduction

Advances in medical imaging technologies, various imaging modalities have been extensively used in the clinical diagnosis and treatment. However, single-modal imaging often fails to capture both structural and functional lesion characteristics, which can reduce diagnostic accuracy and affect clinical decision-making (Azam et al., 2022; Huang et al., 2020). Different imaging modalities provide complementary medical semantic information. For example, Computed Tomography (CT) provides high-resolution anatomical information of bone structures and tissue density, whereas Magnetic Resonance Imaging (MRI) offers exceptional soft-tissue contrast and functional characterization. The integration of CT and MRI enables more accurate lesion localization, tumor delineation, and preoperative planning (Grosu et al., 2005). Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) both capture metabolic and physiological processes. The fusion of SPECT or PET with MRI enhances the precision of functional assessment by combining metabolic information with high-resolution structural details, thereby improving tumor detection, perfusion evaluation, and therapeutic monitoring

* Corresponding author.

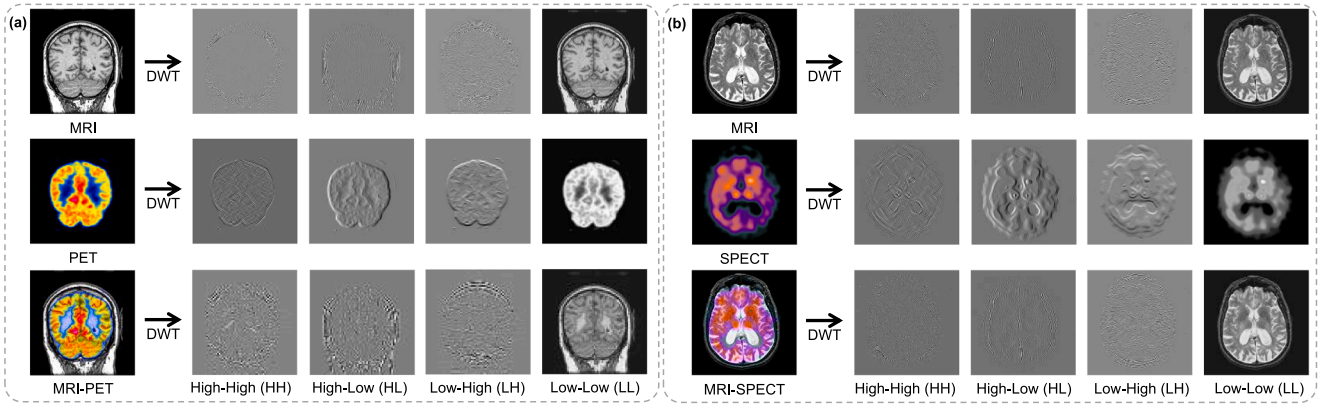

Fig. 1. DWT decomposition results of different MMIF tasks. The DWT effectively separates low-frequency structural information from high-frequency directional details, which provides an intuitive basis for analyzing the complementary relationships between different modalities. LL denotes the low-frequency sub-band preserving overall structures, while LH, HL, and HH represent high-frequency sub-bands corresponding to horizontal, vertical, and diagonal directional details, respectively. (a) MRI-PET fusion task; (b) MRI-SPECT fusion task.

(Jimenez-Mesa et al., 2023). Moreover, combining Ultrasound (US) with Photoacoustic (PA) imaging integrates real-time anatomical observation with high-contrast vascular and molecular information, enabling detailed visualization of microvascular dynamics and tissue functionality (Kim et al., 2024; Park et al., 2024). Multimodal medical image fusion (MMIF) integrates complementary information from different modalities, significantly enhancing diagnostic accuracy and clinical efficiency.

Convolutional neural network (CNN)-based methods exploit their powerful feature extraction and parameter-sharing mechanisms to automatically learn cross-modal complementary information, thereby becoming the efficient approaches for image fusion (Fu et al., 2021; Li & Wu, 2018; Liu et al., 2023a; Wen et al., 2023). Transformer-based fusion methods utilize the self-attention mechanism of Vision Transformers (ViTs) to overcome the locality limitation of convolutional models and demonstrate stronger feature representation and fusion performance in MMIF tasks (Dosovitskiy et al., 2020; Kirillov et al., 2023; Li et al., 2023; Rao et al., 2022). Recently, the Mamba network, based on structured state space models (SSMs), has shown outstanding performance in image processing. Mamba captures long-range dependencies with linear computational and memory complexity, offering higher modeling efficiency than Transformer architectures, which presents strong potential for high-resolution MMIF (Li et al., 2025; Ma et al., 2024).

Most existing DL-based fusion methods primarily focus on extracting and integrating spatial-domain features. In fact, medical images are often affected by noise and artifacts, which compromise the reliability and discriminability of spatial-domain representations (Zhang et al., 2025). Incorporating frequency-domain features offers a more accurate and efficient approach to distinguishing structural information from irrelevant interference. Furthermore, common anatomical structures in medical images, such as blood vessels and nerve fibers, exhibit distinct directional characteristics (Yang & Farsiu, 2023). Therefore, integrating direction-aware modeling into the network can further improve its representational ability. The Discrete Wavelet Transform (DWT), a widely used frequency-domain transformation method, decomposes spatial-domain features into multiple frequency sub-bands, thereby enabling multi-scale frequency representation. Additionally, DWT extracts detail information in the horizontal, vertical, and diagonal directions, exhibiting a certain degree of directional sensitivity and effectively enhancing the representation of structural and textural features. Fig. 1 illustrates the results of applying DWT to source and fused images in MMIF task. The low-frequency sub-band (Low-Low, LL) preserves the overall structural information of the image, whereas the three high-frequency sub-bands (Low-High, LH; High-Low, HL; High-High, HH) emphasize edge and texture details from different directions, which demonstrates the superiority of frequency-domain features in structural fidelity and detail representation. Therefore, incorporating joint frequency-directional modeling can suppress noise and artifacts while enhancing direction-sensitive details such as edges and textures, which improves the model's capacity to represent source image features and enhances the structural fidelity of the fused results.

Although DL-based MMIF methods have achieved satisfactory fusion performance, they still exhibit several limitations: (i) In feature extraction, most existing methods (Liu et al., 2023b; Singh et al., 2009; Srivastava et al., 2025; Wen et al., 2023) primarily focus on spatial-domain features while neglecting spectral-domain information, which leads to a loss of high-frequency details during deep downsampling. Moreover, the absence of effective direction-aware modeling makes it difficult to preserve anatomical structures with distinct orientations, which degrades the structural fidelity of the fused images. (ii) In terms of the fusion mechanism, most methods (Di et al., 2024; He et al., 2025; Tang et al., 2022a; Zhou et al., 2025) employ channel concatenation, weighted fusion, or attention-guided strategies but do not adequately model semantic relationships and cross-modal complementarity, which impairs the discriminability and structural fidelity of the fused images. (iii) Regarding the loss function design, current methods (He et al., 2025; Liu et al., 2023b; Srivastava et al., 2025; Tang et al., 2022a; Zhou et al., 2025) mainly rely on low-level constraints such as the L1 norm, structural similarity, and gradient consistency, but they lack selective guidance based on modality reliability and key structural regions, which makes it difficult to ensure stable and high-quality structural representation in the fused results when noise interference or large modality discrepancies occur.

To address these challenges, we propose a novel frequency-domain and parallel Mamba-based MMIF network, termed FDPMambaFuse. First, we construct a multi-level CNN-Mamba feature extractor, in which CNNs capture shallow spatial information, whereas the proposed DWT-based parallel Mamba module performs deep semantic modeling. In this module, the DWT decomposes spatial features into multi-scale, multi-directional frequency components, while the proposed multi-directional parallel Mamba (MSD-Mamba) models long-range dependencies and enhances direction-sensitive feature representations. This design enables joint characterization of spatial, frequency, and directional information while maintaining efficient global modeling, thereby enhancing the expressive power of heterogeneous modality features. Furthermore, we introduce a multi-stage feature fusion module that integrates cross-modal semantics from shallow to deep levels by combining channel exchange with spatial adaptive attention. This module first promotes shallow semantic interaction through cross-modal channel exchange and then uses an attention mechanism to emphasize critical information, which enables effective feature aggregation and enhancement.

To better guide unsupervised fusion training, we propose a pixel-level uncertainty-aware loss complemented by a gradient-based structural loss, which jointly optimizes fusion quality from both pixel-intensity and structural perspectives. In addition, to address the limitations of existing open-source MMIF datasets in modality diversity and clinical applicability, we construct and release a high-quality multimodal fusion dataset named HIT-MMIF-PAUS. The dataset contains clinically significant PA-US image pairs, which include multiple imaging types and provide more challenging real-world scenarios for evaluating the performance of fusion networks. The main contributions of this work are summarized as follows:

1. We propose FDPMambaFuse, a novel MMIF framework that integrates frequency-domain decomposition with the multi-directional parallel Mamba.
2. We develop a DWT-based parallel Mamba module that jointly captures spatial-frequency-directional information across different modalities.
3. We design a multi-stage feature fusion module that progressively integrates shallow-to-deep cross-modal semantics through channel exchange and spatial adaptive attention.
4. We introduce a pixel-level uncertainty-aware loss combined with a gradient-based structural loss to guide unsupervised fusion training.
5. We construct and release a clinically meaningful PA-US dataset, HIT-MMIF-PAUS, providing diverse and realistic scenarios for MMIF performance evaluation.

The remainder of this paper is organized as follows: Section 2 reviews related work on MMIF. Section 3 details the overall architecture and core components of FDPMambaFuse. Section 4 introduces the experimental setting details, comparative results, and ablation studies. Finally, Section 5 and Section 6 provide the discussion and conclusion of the study.

2. Literature review

2.1. Traditional technique-based MMIF

Traditional technique-based MMIF methods can be broadly categorized into frequency-domain approaches, spatial-domain approaches, and sparse representation (SR)-based approaches. Frequency-domain approaches are typically represented by the DWT and the non-subsampled shearlet transform (NSST). These methods fuse high- and low-frequency components through multi-scale decomposition, which preserves structural information and fine details. However, these methods lack translation invariance, which makes the inverse transformation prone to artifacts. Guihong et al. (2001), Parmar and Kher (2012), Singh et al. (2009), Yin et al. (2018) Spatial-domain approaches operate directly at the pixel level and offer advantages in terms of high efficiency and ease of implementation. For example, strategies based on HSV models or principal component analysis with stationary wavelet transform can highlight the complementarity between structure and function. However, these methods exhibit limited robustness to variations in resolution and complex imaging conditions, making them prone to spatial distortions. Bashir et al. (2019), Li et al. (2021), Stokking et al. (2001) SR-based approaches extract significant features through dictionary learning, and convolutional sparse models leveraging morphological component analysis can improve fusion accuracy. However, their performance depends heavily on dictionary quality and block-partition strategies, which also lead to relatively high computational complexity. Liu et al. (2019), Yang and Li (2009), Zhang et al. (2018) Overall, these methods rely on manually designed rules or shallow-level features, which makes it difficult to achieve end-to-end optimization and capture deep semantic representations. Therefore, they show limited adaptability in complex clinical scenarios.

2.2. Deep learning-based MMIF

In recent years, DL has significantly advanced MMIF by enabling the efficient joint modeling of complementary information from different imaging modalities within end-to-end network frameworks. CNNs have established the classical encode-fuse-decode paradigm for multimodal image fusion. DenseFuse integrates multi-level features via dense connections, whereas U2Fusion and IGNFusion introduce adaptive information preservation mechanisms and cross-scale attention, respectively, to enhance the unsupervised fusion capability in adaptively modeling features across different modalities (Li & Wu, 2018; Wang et al., 2022; Xu et al., 2020). Additionally, F-DARTS enhances structural representation and visual quality by incorporating neural architecture search (Ye et al., 2023). However, CNNs are constrained by local convolutional kernels and have difficulty modeling long-range dependencies and structural correspondences across modalities. Additionally, CNNs lack adaptability to noise variations and modality heterogeneity in medical imaging. Transformer-based methods facilitate global modeling using the self-attention mechanism. SwinFusion explicitly captures both intra- and cross-domain structural associations, improving texture consistency (Ma et al., 2022). Hybrid CNN-Transformer architectures strike a balance between local detail preservation and global relationship modeling, further enhancing fusion performance (Luo et al., 2025; Tang et al., 2022b). Structured State Space Sequence models (S4) efficiently capture long-range dependencies through linear recurrence. Building on this, Mamba introduces a selective mechanism that reduces complexity significantly while preserving strong global modeling capabilities, making it a promising alternative to Transformers (Liu et al., 2024; Zhu et al., 2024). Mambadfuse is the first to apply Mamba modules to multimodal image fusion tasks, achieving excellent fusion results (Li et al., 2024). SAFFusion adopts a Mamba-UNet structure to extract and fuse multi-scale features from source images, effectively leveraging Mamba's strength in long-range dependency modeling to enhance fusion performance (Liu et al., 2025).

2.3. Frequency-domain feature-based MMIF

An increasing number of studies have highlighted the importance of frequency-domain information in MMIF. NSST-CNP employs the NSST for multiscale decomposition and uses coupled neural P systems to improve the preservation of structural and textural details (Satyanarayana & Mohanaiah, 2025). MsgFusion jointly models spatial- and frequency-domain representations under medical semantic guidance, which promotes cross-modal semantic complementarity (Wen et al., 2023). MMIF-INet integrates the DWT with invertible neural networks, which enables information-lossless and multiscale frequency fusion (He et al., 2025). SAFFusion introduces a saliency-aware frequency fusion strategy and leverages a Mamba-UNet architecture, which strengthens detail representation in clinically critical regions (Liu et al., 2025). DWM-Fusion combines discrete wavelet convolution with Mamba, which strikes a better balance between fusion performance and computational efficiency (Fan et al., 2025). SFDKAN performs spatial-frequency dual-domain fusion with Kolmogorov-Arnold networks, which improves high-frequency detail representation and global feature association (Lin et al., 2025). These studies collectively indicate that frequency-domain modeling has become a key direction for improving MMIF performance. However, existing frequency-domain fusion methods generally lack explicit modeling of directional characteristics, which limits their ability to maintain directional consistency and edge continuity of anatomical structures across modalities.

3. Methodology

The overall architecture of the proposed FDPMambaFuse is shown in Fig. 2. It consists of three main steps: multi-level feature extraction,

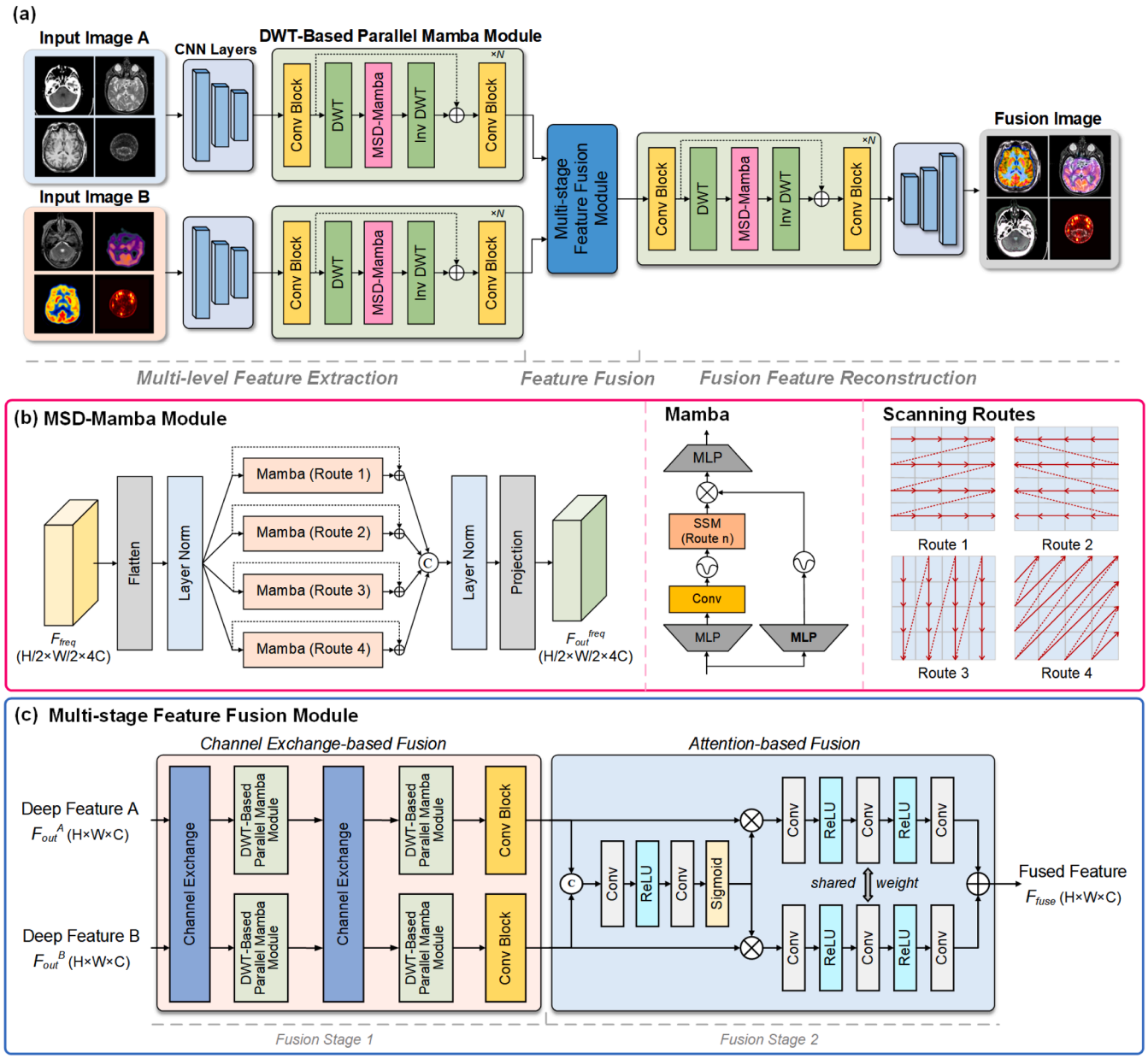

Fig. 2. The overview of our FDPMambaFuse. FDPMambaFuse mainly consists of three stages: multi-level feature extraction, multi-stage feature fusion, and fused feature reconstruction. (a) Structure of FDPMambaFuse. (b) Structure of the MSD-Mamba module, Mamba structure, and schematic of the scanning directions. (c) Structure of the multi-stage feature fusion module.

multi-stage feature fusion, and fusion feature reconstruction. The network receives a pair of images from different modalities as input. For grayscale modalities such as CT, MRI, and US, the images are directly used without additional preprocessing. But for pseudo-color modalities like PET, SPECT, and PA are usually represented in RGB format. These RGB images are first converted to the YCbCr color space, from which the luminance (Y) channel is extracted and used in the fusion process to preserve structural and intensity information. After fusion, the resulting Y channel is combined with the original chrominance components (Cb and Cr), and the image is converted back to RGB to produce the final fused result. In this process, only the luminance (Y) channel is used instead of chromaticity information because MRI and US are single-channel grayscale images that contain only luminance information and cannot provide color components. Therefore, color information is entirely provided by pseudo-color modes, such as PET, SPECT, and PA, eliminating the need for chromaticity components from grayscale images, which effectively preserves both structural and color information.

3.1. Multi-level feature extractor

To effectively extract spatial, frequency-domain, and directional features from multimodal images, we design a multi-level feature extractor. This extractor consists of two levels: the shallow feature extraction module, which focuses on extracting local textures and edge features in the spatial domain, and the deep feature extraction module, which extracts frequency-directional features and captures multi-scale high-frequency details and directional structural information.

3.1.1. CNN-based shallow-level feature extraction

First, the source images $I_A \in \mathbb{R}^{H\times W\times 1}$ and $I_B \in \mathbb{R}^{H\times W\times 1}$ are input into a shallow feature extractor consisting of three convolutional blocks. This extractor captures rich local texture and edge details from the multimodal inputs, generating the corresponding feature maps $F_A \in \mathbb{R}^{H\times W\times C}$ and $F_B \in \mathbb{R}^{H\times W\times C}$. Each convolutional block consists of a 3×3 convolutional layer followed by a LeakyReLU activation function. The convolutional layers extract fine-grained spatial features, while the nonlinear activation enhances the network's representational capacity. The shallow feature extractor preserves rich local structural information in the spatial domain, providing a foundation for subsequent high-level semantic modeling and cross-modal fusion.

3.1.2. Mamba-based high-level feature extraction

Key anatomical structures in medical images exhibit distinct scale- and direction-specific characteristics in the frequency domain, making it challenging for spatial-domain modeling to balance global semantic consistency with local edge fidelity. The DWT explicitly separates low-frequency structures and multi-directional high-frequency details in a single decomposition while preserving spatial locality, making it particularly suitable for generating compact and discriminative representations in medical images. While frequency decomposition provides structured directional subband representations, modeling long-range dependencies across directions and scales requires additional contextual reasoning mechanisms, especially in anatomically complex regions. Based on these considerations, after obtaining the shallow features, we design a high-level feature extractor to enhance the ability to model global structures and high-level semantic relationships in multimodal images. This extractor consists of N DWT-based parallel Mamba modules ϕ_n. Specifically, a discrete wavelet decomposition is applied to the shallow features $F \in \mathbb{R}^{H\times W\times C}$, performing a single-level 2D DWT that produces one low-frequency and three directional high-frequency components: $Y_{LL}, Y_{LH}, Y_{HL}, Y_{HH} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$. We adopt a Daubechies-4 wavelet with fixed filter parameters, which provides favorable regularity and stability for medical image representation. The process is defined as:

$$(Y_{LL}, Y_{LH}, Y_{HL}, Y_{HH}) = \mathrm{DWT}(F) \tag{1}$$

where $\mathrm{DWT}(\cdot)$ denotes the DWT operation. The frequency-enhanced feature $F_{freq} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$ are obtained by:

$$F_{freq} = \mathrm{Concat}(Y_{LL}, Y_{LH}, Y_{HL}, Y_{HH}) \tag{2}$$

where $\mathrm{Concat}(\cdot)$ denotes the concatenation operation. After obtaining the frequency-enhanced feature F_{freq}, the representation contains rich multi-directional and cross-scale information that is difficult to capture using a single-path state modeling strategy. Through frequency decomposition, spatial features are explicitly separated into subbands with distinct directional and scale characteristics, providing a structured representation for subsequent modeling.

Based on this structured frequency representation, we design MSD-Mamba to perform state-space modeling along multiple propagation paths. Each path is defined by a specific sequence reordering strategy, which induces a distinct recursive dependency pattern over the frequency-organized features, thereby altering the state transition topology across frequency-structured tokens. By modeling these complementary propagation paths in parallel, the proposed module establishes a unified frequency-state modeling framework. The overall structure of MSD-Mamba is illustrated in Fig. 2(b).

Specifically, F_{freq} is first flattened into a 2D sequence and then normalized:

$$X = \mathrm{LN}(\mathrm{Flatten}(F_{freq})) \tag{3}$$

where $X \in \mathbb{R}^{\frac{H\cdot W}{4}\times 4C}$ denotes the normalized and flattened features. $\mathrm{Flatten}(\cdot)$ indicates the flattening operation, and $\mathrm{LN}(\cdot)$ denotes layer normalization.

Next, the normalized feature sequence is divided into four parallel subsequences, denoted as $X = [X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}]$, where $X^{(i)} \in \mathbb{R}^{\frac{HW}{4}\times C}$ for $i = 1, 2, 3, 4$. Different sequence reordering operators are then applied to construct four input sequences with distinct arrangements:

$$\hat{X}^{(i)} = \mathcal{R}_i(X^{(i)}) \in \mathbb{R}^{\frac{HW}{4}\times C}, \quad i = 1, 2, 3, 4. \tag{4}$$

The reordering operators are defined as:

$$\hat{X}^{(1)} = X^{(1)} \tag{5}$$

$$\hat{X}^{(2)} = \mathrm{Reverse}\big(X^{(2)}\big) \tag{6}$$

$$\hat{X}^{(3)} = \mathrm{Permute}_{col}\big(X^{(3)}\big) \tag{7}$$

$$\hat{X}^{(4)} = \mathrm{Permute}_{diag}\big(X^{(4)}\big) \tag{8}$$

where, $\mathrm{Reverse}(\cdot)$ denotes sequence reversal along the temporal dimension, while $\mathrm{Permute}_{col}(\cdot)$ and $\mathrm{Permute}_{diag}(\cdot)$ represent column-wise and diagonal traversal reordering strategies, respectively. These operators induce complementary state transition patterns along horizontal, vertical, and diagonal orientations, allowing the SSM to capture anisotropic long-range dependencies that are difficult to model using a single sequential ordering.

Each reordered subsequence $\hat{X}^{(i)}$ is independently processed by a corresponding Mamba branch. Specifically, two multilayer perceptrons are used to generate the main feature branch and the gating vector:

$$x^{(i)} = \mathrm{MLP}_x(\hat{X}^{(i)}), \quad z^{(i)} = \mathrm{MLP}_z(\hat{X}^{(i)}). \tag{9}$$

where $x^{(i)}, z^{(i)} \in \mathbb{R}^{\frac{HW}{4}\times C}$. $\mathrm{MLP}(\cdot)$ denotes MLP. The feature $x^{(i)}$ is further enhanced through local modeling using the convolutional operation $\mathrm{Conv}(\cdot)$ followed by the $\mathrm{SiLU}(\cdot)$ activation:

$$\tilde{x}^{(i)} = \mathrm{SiLU}(\mathrm{Conv}(x^{(i)})), \tag{10}$$

where $\tilde{x}^{(i)} \in \mathbb{R}^{\frac{HW}{4}\times C}$ denotes the locally augmented trunk path feature. Subsequently, $\tilde{x}^{(i)}$ is input into the state space model (SSM), which performs recursive dependency modeling under the topology induced by the corresponding reordering operator, yielding the direction-aware feature output $f^{(i)} \in \mathbb{R}^{\frac{H\cdot W}{4}\times C'}$:

$$f^{(i)} = \mathrm{SSM}(\tilde{x}^{(i)}) \tag{11}$$

where $\mathrm{SSM}(\cdot)$ performs dynamic parameter generation and sequence recursion required for state modeling. The feature $f^{(i)}$ is fused with gating vector, and the final direction-aware output $Y^{(i)} \in \mathbb{R}^{\frac{H\cdot W}{4}\times C}$ is obtained through the residual connection and mapping layer:

$$Y^{(i)} = \mathrm{MLP}_f\left(f^{(i)} \odot \mathrm{SiLU}(z^{(i)})\right) + X^{(i)} \tag{12}$$

Finally, the modeling results from the four directions are concatenated and restored to the original spatial shape to form the multi-directional structure-aware enhanced feature $F_{out}^{freq} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$:

$$F_{out}^{freq} = \mathrm{Pro}(\mathrm{LN}(\mathrm{Concat}(Y^{(1)}, Y^{(2)}, Y^{(3)}, Y^{(4)}))) \tag{13}$$

where $\mathrm{Pro}(\cdot)$ denotes the projection operation.

The four subband features are obtained by splitting F_{out}^{freq} along the channel dimension and are reconstructed into spatial-domain representations using the inverse discrete wavelet transform (IDWT). Through IDWT-based reconstruction and residual fusion, the frequency-domain dependency modeling results are continuously injected back into the spatial feature space, forming an explicit feedback mechanism for subsequent layers. The output of the DWT-based parallel Mamba module is generated through residual concatenation followed by convolution. By stacking N such modules, the deep feature representation $F_{out} \in \mathbb{R}^{H\times W\times C}$ is progressively extracted:

$$F_{out} = \phi_N \circ \cdots \circ \phi_2 \circ \phi_1(F) \tag{14}$$

where ϕ_N represents the operation of the DWT-based parallel Mamba module.

3.2. Multi-stage feature fusion

Given that the multimodal medical image features extracted in the earlier stage still exhibit substantial discrepancies in terms of resolution, contrast, noise distribution, and structural representation, and considering that anatomical relevance varies across different spatial regions, we propose a multi-stage feature fusion module to achieve efficient and adaptive cross-modal fusion, as illustrated in Fig. 2(c). The first stage performs cross-modal feature topology reorganization and embeds directional structural priors into spatial representations. The second stage applies spatially adaptive modulation via a dual-branch attention mechanism on the reorganized features.

To address the differences in noise levels and detail representation between structural and functional modalities while avoiding mutual interference between their information. In the first stage, deep features $F_{out}^{A}, F_{out}^{B} \in \mathbb{R}^{H\times W\times C}$ from both modalities obtained by the feature extraction extractors are partially exchanged in the channel dimension (Fang et al., 2023). Specifically, features from the even-indexed channels of F_{out}^{A} are exchanged with the corresponding channels in F_{out}^{B}, enabling lightweight interaction and fusion of multimodal information without additional computational overhead. This operation explicitly reorganizes the cross-modal information flow in feature space, forming interleaved modality representations that facilitate subsequent structured modeling.

Subsequently, the reorganized features are processed by the proposed DWT-based parallel Mamba module. By decomposing the features into wavelet subbands and performing multi-directional state-space modeling, the module captures directional and cross-scale dependency patterns. These dependencies are then embedded into the spatial feature space through IDWT-based reconstruction, enabling the reconstructed spatial features to encode direction-aware structural relationships. This process is repeated, and the initial fused features $F_{f1}^{A}, F_{f1}^{B} \in \mathbb{R}^{H\times W\times C}$ are obtained through a convolutional module. The resulting spatial features already encode directional structural dependencies and serve as topology-aware representations for the subsequent fusion stage.

Since different spatial regions exhibit modality-specific advantages in structural representation or functional response, in the second stage we introduce a lightweight dual-branch spatial attention mechanism to perform pixel-wise adaptive modulation. The attention mechanism operates on the topology-reorganized and direction-embedded spatial features obtained from the first stage, adaptively adjusting modality contributions across anatomical regions. Specifically, the initial fused feature pair F_{f1}^{A} and F_{f1}^{B} is first concatenated to form a joint spatial representation that integrates the structurally enriched features from both modalities, and is then processed through convolution, ReLU activation, and Sigmoid normalization to generate two spatial attention maps $A_1, A_2 \in \mathbb{R}^{H\times W\times C}$:

$$A_1, A_2 = \mathrm{Sigmoid}(\mathrm{Conv}(\mathrm{ReLU}(\mathrm{Conv}(\mathrm{Concat}(F_{f1}^{A}, F_{f1}^{B})))))) \tag{15}$$

where $\mathrm{Sigmoid}(\cdot)$ represents the Sigmoid activation function. The resulting attention maps A_1 and A_2 are then element-wise multiplied with F_{f1}^{A} and F_{f1}^{B}, respectively, to adaptively weight the feature contributions of each modality across different anatomical regions.

The weighted features are then individually fed into a weight-sharing feature enhancement module composed of convolutional blocks to improve the consistency and synergy of structural representations across modalities. Finally, the enhanced features from both modalities are fused through element-wise addition to obtain the final fused feature $F_{fuse} \in \mathbb{R}^{H\times W\times C}$ for subsequent reconstruction.

3.3. Fusion feature reconstruction

The fusion feature reconstruction module adopts a structure symmetric to the feature extraction stage. It comprises multiple DWT-based parallel Mamba modules and a three-layer convolutional block, jointly preserving key information from the original modalities and enabling high-quality reconstruction of the fused features. For input pairs composed of two grayscale modalities, the reconstructed fused feature directly serves as the final fusion result. However, when the input pair involves a pseudo-color modality such as PET, SPECT, or PA in RGB format, only the luminance channel Y of the pseudo-color modality participates in the fusion process. After reconstructing the fused Y-channel, it is recombined with the original chrominance components Cb and Cr retained from preprocessing, and the fused YCbCr image is finally converted back to RGB space to produce the final pseudo-color fused result.

3.4. Loss function

As MMIF is an unsupervised learning framework, the loss function plays a pivotal role in steering the network toward producing reliable and clinically meaningful fusion results. We propose a hybrid loss function that integrates a pixel-level uncertainty-guided loss $\mathcal{L}_{un}$ and an image gradient-based loss $\mathcal{L}_{grad}$. And the complete loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{un} + \lambda_2 \mathcal{L}_{grad} \tag{16}$$

where $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ are weighting parameters. The overall loss function imposes complementary constraints at two levels, global semantic information selection and local detail preservation, to jointly enhance the fusion quality.

The pixel-level uncertainty-guided loss $\mathcal{L}_{un}$ models cross-modal reliability variations to construct adaptive supervision for multimodal fusion. In multimodal medical imaging, pixel intensities at the same spatial location often exhibit modality-dependent characteristics, making direct consistency constraints suboptimal. By estimating local cross-modal reliability, the proposed loss enables spatially adaptive supervision that reflects modality complementarity. To this end, we first compute the pixel-level uncertainty map U between the input images I_A and I_B:

$$U = \left| I_A - I_B \right| \tag{17}$$

where U reflects the cross-modal consistency at each pixel location, with lower uncertainty indicating more reliable information at that position. Although U is computed using pixel-wise intensity differences, it functions as a cross-modal reliability prior that guides adaptive supervision construction.

Based on the uncertainty map, confidence-weighting coefficients ω_A and ω_B are generated to dynamically construct a pseudo-supervision target rather than reweighting an existing one. These coefficients are then used to generate the pseudo-fusion target map as follows:

$$\begin{aligned} T &= \omega_A I_A + \omega_B I_B \\ \omega_A &= \exp(-U), \quad \omega_B = 1 - \omega_A \end{aligned} \tag{18}$$

This generation strategy emphasizes the structural modality in cross-modality consistent regions and enhances the functional modality in regions with modality discrepancies, allowing the pseudo-labels to naturally reflect the complementary relationship between structural and functional information. Then $\mathcal{L}_{un}$ is calculated as:

$$\mathcal{L}_{un} = \frac{1}{N_0} \sum_{i=1}^{N_0} U_i \cdot (F_i - T_i)^2 \tag{19}$$

where F_i denotes the ith pixel position of the fused image, and $N_0 = H \times W$ represents the total number of pixels in the image. During optimization, the uncertainty map U further modulates the contribution of each spatial location to the overall objective, coupling supervision construction with confidence-aware regulation. Regions with higher uncertainty are assigned stronger corrective influence, while more consistent regions are guided with smoother constraints. This adaptive regulation stabilizes learning in cross-modal conflict areas and improves structural fidelity in the fused results.

To better preserve structural details and edge information from the input modalities, we additionally incorporate an image gradient loss to

constrain the similarity between the fused image and the source images at the edge and texture levels. First, the Sobel operator ∇ is applied to extract image gradient information in both horizontal and vertical directions. Then, the differences in gradient space between the fused image F and the two input modality images I_A and I_B. $\mathcal{L}_{grad}$ is calculated as:

$$\mathcal{L}_{grad} = \frac{1}{N_0}\left\|\nabla F - \max(\nabla I_A, \nabla I_B)\right\|_2 \tag{20}$$

where $\|\cdot\|_2$ denotes the L2 norm, and $\max(\cdot)$ represents the maximum value computation.

4. Experimental results

4.1. Experimental dataset and setup

For our experiments, we utilized the publicly available brain atlas dataset from Harvard University (https://www.med.harvard.edu/AANLIB/home.html), and additionally constructed and open-sourced a multimodal dataset, HIT-MMIF-PAUS, specifically for the PA-US image fusion task.

The brain atlas dataset includes CT, MRI, SPECT, and PET modalities, encompassing a wide range of pathological conditions, including cerebrovascular diseases, brain tumors, degenerative disorders, inflammatory conditions, and normal brains. The dataset provides images in multiple anatomical planes (axial, sagittal, and coronal) with a uniform resolution of 256×256. All images are spatially aligned and preprocessed, making them widely used in multimodal medical image fusion research.

The HIT-MMIF-PAUS dataset includes various imaging targets, such as subcutaneous mouse tumors, human hand skin, human fingers, and mouse rectum, commonly used in medical tasks such as tumor detection and microvascular visualization. The image pairs are acquired using three different types of PA-US imaging systems: i) The linear-array imaging system utilizes the commercial S-Sharp Prodigy system (China), combined with a 532nm pulsed laser (SpitLight, Innolas, Germany) and a 128-element handheld ultrasound probe, producing original images of size 256×256. This system is suitable for high-resolution imaging of superficial and soft tissues, providing fine anatomical details, particularly for tumor and skin lesion detection. ii) The ring-array imaging system uses a custom fiber optic and a 980nm pulsed laser (GmbH, Innolas, Germany), along with a ring-shaped ultrasound transducer array, producing original images of size 512×512. This system enables deeper tissue imaging, suitable for large-area structural imaging and vascular visualization, providing high-contrast functional images, ideal for observing microvessels and deeper tissues. iii) The endoscopic imaging system consists of a 532nm pulsed laser (MIRON-DPSSL-PL, MIRON, Germany), an ultrasound transmission/reception unit (5073PR, Olympus, Japan), and a photoelectric slip ring (PT1-Z8, Thorlabs, Germany), producing original images of size 512×512. This system features endoscopic imaging capabilities, suitable for high-resolution imaging in confined spaces, particularly useful for observing complex structures, such as the gastrointestinal tract. All imaging systems combine both photoacoustic and ultrasound imaging functions, ensuring consistent spatial resolution across modalities and precise image registration. These three imaging approaches differ in their viewing angles and imaging depths, with the selected scan target areas being clinically representative and challenging, ensuring the dataset's clinical relevance and applicability. To facilitate uniform training and testing of the network, we standardized the image pair size to 256×256. Additionally, to support downstream segmentation tasks for MMIF fusion, we created a linear-array PA-US tumor segmentation dataset. The labels for this dataset were generated by clinical professionals based on a comprehensive analysis of both PA and US imaging results.

A total of 300 image pairs are selected for training and testing, covering four modality combinations: MRI-CT, MRI-PET, MRI-SPECT, and PA-US. The training set contains 40 image pairs for each modality combination, while the testing set consists of 30 MRI-CT, 60 MRI-PET, 30 MRI-SPECT, and 20 PA-US image pairs.

The experiments are conducted on a system equipped with an Intel(R) Xeon(R) Platinum 8370C CPU @ 2.80 GHz and NVIDIA GeForce RTX 4090 GPUs. The implementation is based on the PyTorch framework. The model is optimized using the AdamW optimizer. An initial learning rate of 0.0001 is adopted, and a cosine annealing strategy is employed to dynamically adjust the learning rate during training. All input images are resized to 256×256, and the batch size is set to 6. The model is trained for 300 epochs. The parameter N in the DWT-based Parallel Mamba module is set to 9.

4.2. Comparison methods and evaluation metrics

To validate the effectiveness of the proposed FDPMambaFuse, we compared it with two traditional fusion methods, including CSMCA (2019) (Liu et al., 2019) and NSST-PAPCNN (2019) (Yin et al., 2018), and eight DL-based methods, including IFCNN (2020) (Zhang et al., 2020), U2Fusion (2022) (Xu et al., 2020), SwinFusion (2022) (Ma et al., 2022), MM-Net (2024) (Liu et al., 2024), SAFFusion (2025) (Liu et al., 2025), DWMFusion (2025) (Fan et al., 2025), Mambadfuse (2024) (Li et al., 2024), and MMIF-INet (2025) (He et al., 2025). To ensure the fairness of the comparison, all methods were implemented under the same experimental settings and preprocessing procedures, and were configured strictly according to the optimal hyperparameters reported in their original publications.

To comprehensively and quantitatively evaluate fusion performance, we selected three categories of evaluation metrics: (i) Information retention metrics assess whether the fused image effectively preserves key information from the source images. Entropy (EN) reflects the overall information richness of the image, with higher values indicating greater information content. Mutual Information (MI) quantifies the amount of shared information between the fused image and the source images. Quality Mutual Information (Q_{MI}) incorporates perceptual quality weights into mutual information to comprehensively evaluate the effectiveness of information transfer and perceptual relevance. (ii) Clarity and detail retention metrics evaluate edge sharpness and detail preservation in the fused image. Standard Deviation (SD) describes the dispersion of grayscale values, which is associated with image contrast and visual depth. Average Gradient (AG) quantifies edge sharpness, reflecting the clarity of image details. (iii) Structural fidelity and perceptual quality metrics evaluate structural consistency and perceived image quality from a human visual perspective. Quality Assessment Based on Fusion Features (Q_{ABF}) captures the consistency of edges and local structural information. Structural Similarity Index ($SSIM$) evaluates the similarity between the fused and source images in luminance, contrast, and structure. Visual Information Fidelity ($VIFF$) assesses perceptual quality based on a multi-scale human visual system model.

4.3. Comparative experimental results

4.3.1. Brain atlas image fusion

Fig. 3 presents the qualitative fusion results of all methods on the open-source brain atlas dataset, covering three modality combinations of CT-MRI, MRI-PET and MRI-SPECT, with three fusion examples provided for each combination. To better compare fusion details across methods, representative structural regions were locally enlarged for analysis. Meanwhile, Table 1 presents the quantitative results of all methods across multiple evaluation metrics, comprehensively reflecting performance differences in information retention, texture clarity, and structural fidelity. Fig. 5 further illustrates the average rankings of the methods across different tasks using a line chart.

In the CT-MRI fusion task, FDPMambaFuse demonstrates significantly superior overall performance compared with all SOTA methods. Quantitative results show that it achieves the highest scores across multiple key metrics, including MI of 1.2671, Q_{MI} of 0.7132, SD of 92.2379,

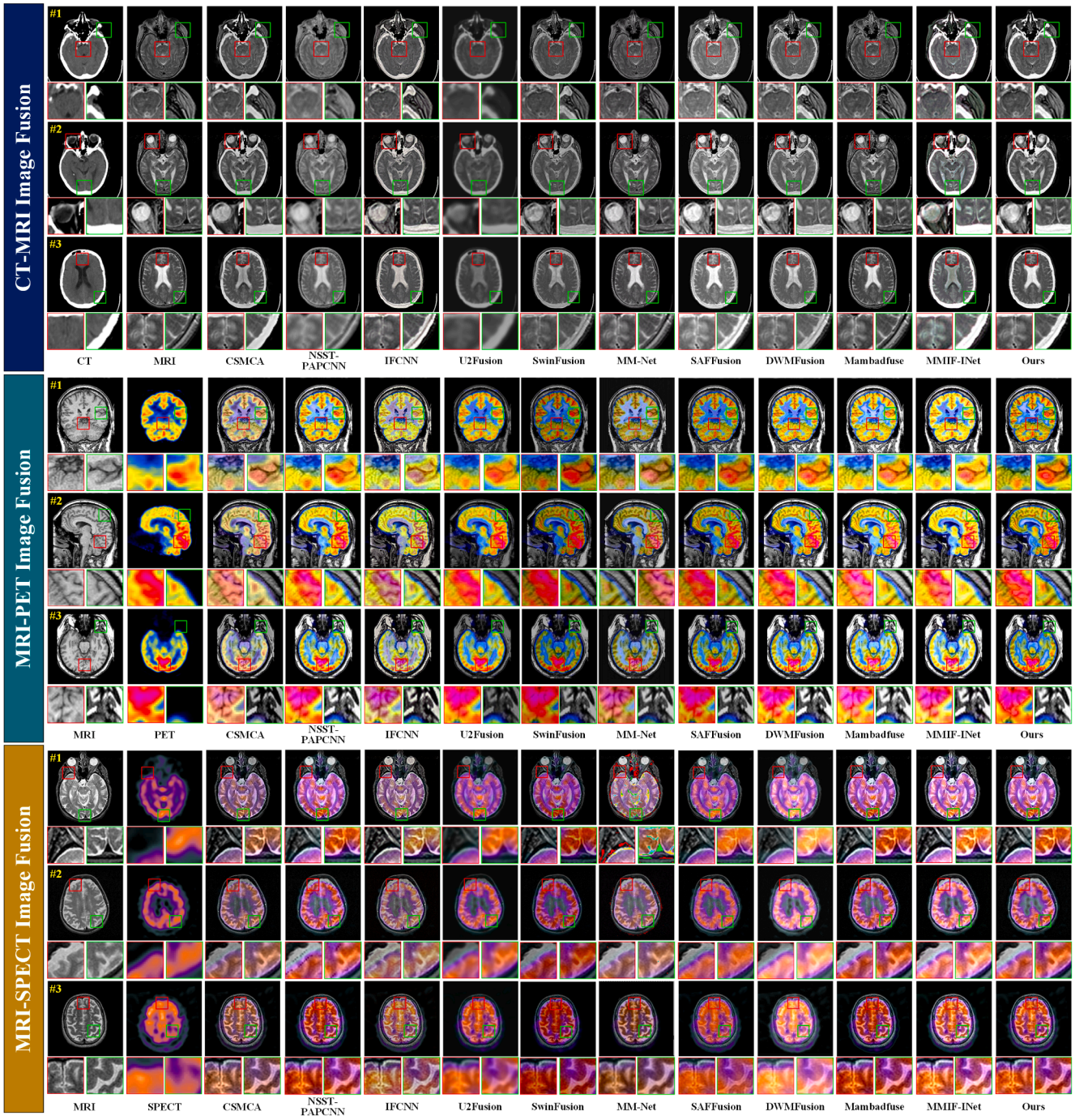

Fig. 3. Qualitative comparison of the proposed method with ten SOTA methods on brain atlas image fusion dataset. From left to right: Two source images, fusion results of CSMCA, NSST-PAPCNN, IFCNN, U2Fusion, SwinFusion, MM-Net, SAFFusion, DWMFusion, Mambadfuse, MMIF-INet, and the proposed method. For better comparison, two local regions are enlarged as close-ups in each image.

Q_{ABF} of 0.6206, $SSIM$ of 0.7862, and $VIFF$ of 0.6134, indicating clear advantages in structural detail preservation, information complementarity, and perceptual quality. Qualitative results further confirm that FDPMambaFuse more effectively retains the color of dense bone regions in enlarged CT areas while preserving MRI soft tissue textures, resulting in higher fusion quality. Compared to other SOTA methods, FDPMambaFuse demonstrates superior performance. For example, SAFFusion, which also relies on frequency-domain features, achieves an MI of 1.1602 and a Q_{MI} of 0.6342. Although SAFFusion performs well in information complementarity, it still lags behind FDPMambaFuse in detail preservation and perceptual quality. DWMFusion achieves an MI of 1.2542, but its Q_{MI} is only 0.5189, and its structural detail preservation is slightly inferior to that of FDPMambaFuse. MambaFuse, based on the Mamba model, performs similarly to FDPMambaFuse, with a Q_{MI} of 0.7052, but its SD is only 56.8507, and it still exhibits shortcomings in detail preservation and information complementarity. MMIF-INet exhibits a high $SSIM$ of 0.7519, indicating strong perceptual quality, but its Q_{MI} is 0.5243 and its $VIFF$ is 0.5919, highlighting limitations in detail preservation and cross-modal consistency. These comparison results demonstrate that FDPMambaFuse not only better preserves structural

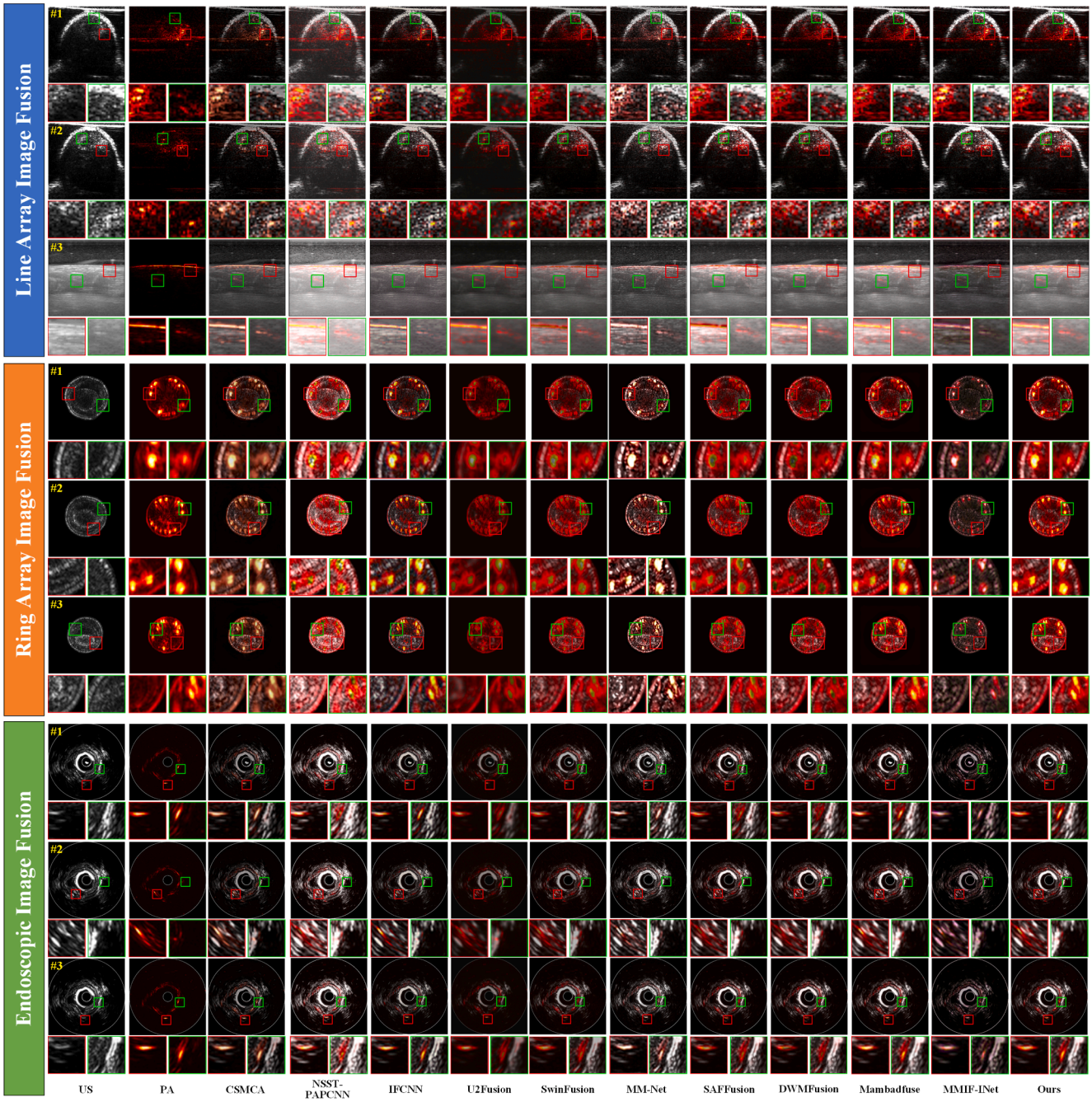

Fig. 4. Qualitative comparison of the proposed method with ten SOTA methods on HIT-MMIF-PAUS image fusion dataset. From left to right: Two source images, fusion results of CSMCA, NSST-PAPCNN, IFCNN, U2Fusion, SwinFusion, MM-Net, SAFFusion, DWMFusion, Mambadfuse, MMIF-INet, and the proposed method. For better comparison, two local regions are enlarged as close-ups in each image.

details and brightness information in CT and MRI fusion, but also outperforms current SOTA methods across multiple key metrics, proving its superiority in MMIF.

In the MRI-PET and MRI-SPECT fusion tasks, FDPMambaFuse also outperforms all SOTA methods in overall performance. In the MRI-PET task, it achieves the highest scores across multiple key metrics, including EN of 6.0834, MI of 1.4775, SD of 91.3421, AG of 14.9363, and $VIFF$ of 0.7316. In the MRI-SPECT task, it likewise leads with MI of 1.5287, SD of 72.8656, AG of 8.2050, and $VIFF$ of 0.8522, while achieving a competitive $SSIM$ of 0.5425. Qualitative results further confirm these advantages. FDPMambaFuse achieves smoother cross-modal structural transitions, reducing blurring and artifacts. It enhances metabolic signal contrast in PET and SPECT functional regions while clearly preserving anatomical boundaries from MRI, thereby achieving an optimal balance between functional information and texture details. In contrast, other methods exhibit varying degrees of limitations. CSMCA and IFCNN fail to effectively retain metabolic features from functional modalities, resulting in insufficient luminance and blurred functional regions in the fused images. U2Fusion, SwinFusion, SAFFusion, DWMFusion, Mambadfuse, and MMIF-INet, although showing improvements in some metrics, still exhibit deficiencies in detail preservation and boundary clarity. U2Fusion in the MRI-PET task achieves an EN of 5.7181, a Q_{ABF} of 0.5812, but its AG is only 7.6140, and $SSIM$ is 0.5397, resulting in texture degradation and blurred anatomical structures. SwinFusion in

Table 1
Quantitative results on the Brain Atlas and HIT-MMIF-PAUS datasets. The best score is marked in **bold**, and the second-best score is underlined.

Dataset	Method	$EN\uparrow$	$MI\uparrow$	$Q_{MI}\uparrow$	$SD\uparrow$	$AG\uparrow$	$Q_{ABF}\uparrow$	$SSIM\uparrow$	$VIFF\uparrow$
CT-MRI	CSMCA (2019)	4.2285	0.8221	0.4723	71.6612	7.2071	0.5756	0.6107	0.4162
	NSST-PAPCNN (2019)	4.3131	0.9057	0.4897	88.5524	8.2332	0.5911	0.5920	0.5382
	IFCNN (2020)	4.2447	0.9027	0.5161	77.9783	**9.1170**	0.5258	0.6229	0.4879
	U2Fusion (2022)	**4.6828**	0.7960	0.4293	50.1812	2.6893	0.4919	0.6324	0.2307
	SwinFusion (2022)	4.4730	1.0274	0.5728	61.6637	4.5956	0.5258	0.6412	0.4753
	MM-Net (2024)	4.2049	1.1979	0.6697	53.6717	5.8962	0.4164	0.7352	0.5553
	SAFFusion (2025)	4.4314	1.1602	0.6342	91.3423	8.2454	0.5933	0.7342	0.5838
	DWMFusion (2025)	4.5132	1.2542	0.5189	91.4525	8.2353	0.6142	0.7454	0.5672
	Mambadfuse (2024)	4.2280	1.2421	0.7052	56.8507	8.9025	0.5986	0.7115	0.6053
	MMIF-INet (2025)	4.2956	0.9202	0.5243	92.2262	8.1185	0.5700	0.7519	0.5919
	Ours	4.6500	**1.2671**	**0.7132**	**92.2379**	8.8579	**0.6206**	**0.7862**	**0.6134**
MRI-PET	CSMCA (2019)	4.8846	0.8400	0.3938	72.8023	11.8342	0.5713	0.6134	0.5266
	NSST-PAPCNN (2019)	5.6914	0.9871	0.4203	87.7785	13.9413	0.5587	0.6390	0.6142
	IFCNN (2020)	5.2565	0.9261	0.4143	79.1981	14.0938	0.5718	0.6471	0.5971
	U2Fusion (2022)	5.7181	1.0393	0.4397	67.1216	7.6140	**0.5812**	0.5397	0.4373
	SwinFusion (2022)	5.1935	1.1250	0.5610	85.4144	13.4357	0.5233	0.6543	0.5181
	MM-Net (2024)	5.4595	0.9784	0.4448	80.0926	11.8719	0.3704	0.6513	0.4286
	SAFFusion (2025)	5.4367	1.3854	0.5645	89.4532	13.9867	0.5038	0.6472	0.7081
	DWMFusion (2025)	5.5872	1.4016	0.4187	91.0892	14.2432	0.5224	0.6574	0.7287
	Mambadfuse (2024)	5.3147	1.3391	**0.5862**	84.2205	13.9003	0.5269	**0.6608**	0.5130
	MMIF-INet (2025)	5.6262	1.4255	0.4204	91.1869	14.2620	0.5219	0.6403	0.7106
	Ours	**6.0834**	**1.4775**	0.5834	**91.3421**	**14.9363**	0.5246	0.6586	**0.7316**
MRI-SPECT	CSMCA (2019)	4.4881	0.9330	0.4817	53.6137	6.7754	0.5012	0.5121	0.5320
	NSST-PAPCNN (2019)	4.8972	1.0632	0.5188	66.0693	7.3707	0.5758	0.5234	0.6469
	IFCNN (2020)	4.5455	0.9638	0.4941	56.4844	7.5695	0.5843	0.5365	0.5973
	U2Fusion (2022)	4.5559	0.8770	0.4520	51.9570	2.7023	0.5262	0.3692	0.3632
	SwinFusion (2022)	4.2563	1.3256	0.6956	59.8798	5.9542	0.3415	0.4914	0.4536
	MM-Net (2024)	4.4704	1.3528	0.6922	59.6864	6.0114	0.2994	0.4978	0.4430
	SAFFusion (2025)	4.4982	1.3645	0.5245	69.7653	7.2582	0.5722	0.5309	0.8296
	DWMFusion (2025)	4.7431	1.4142	**0.7384**	72.1487	7.8864	0.5845	0.5362	0.8432
	Mambadfuse (2024)	4.5187	1.4215	0.7211	62.9224	7.2564	0.5915	0.5045	0.5346
	MMIF-INet (2025)	4.8132	1.0396	0.5147	72.6813	7.8974	0.5731	0.5298	0.8348
	Ours	**4.9369**	**1.5287**	0.7284	**72.8656**	**8.2050**	**0.5941**	**0.5425**	**0.8522**
PA-US	CSMCA (2019)	3.3646	0.4482	0.2168	42.5739	6.0895	0.7530	0.4700	0.5803
	NSST-PAPCNN (2019)	3.4627	0.6098	0.3752	57.0259	6.1368	**0.7725**	0.4644	0.8076
	IFCNN (2020)	3.6528	0.6025	0.3609	52.9080	7.1047	0.7282	0.5334	0.6626
	U2Fusion (2022)	2.1423	0.4352	0.3419	49.2438	5.3647	0.3072	0.3655	0.6252
	SwinFusion (2022)	2.8762	0.7164	0.3372	61.5767	5.5723	0.6832	0.5428	0.6745
	MM-Net (2024)	2.1696	0.7008	0.2811	21.2634	2.3746	0.7637	0.2699	0.4033
	SAFFusion (2025)	3.9145	0.6742	0.3682	64.8740	6.8549	0.6502	0.5584	0.7846
	DWMFusion (2025)	3.9825	0.8645	0.3642	64.7600	6.9045	0.6465	0.5642	0.8145
	Mambadfuse (2024)	3.9522	0.8773	0.3442	64.6255	6.0785	0.7357	0.5630	0.7737
	MMIF-INet (2025)	3.8843	0.6472	0.3669	64.9388	6.8915	0.6510	0.5569	0.7939
	Ours	**4.4448**	**1.0659**	**0.3920**	**67.2982**	**7.3720**	0.7674	**0.5798**	**0.8245**

the MRI-SPECT task achieves an $SSIM$ of 0.6543, showing some edge preservation, but still has limitations in retaining details and functional regions. SAFFusion and DWMFusion show improvements in luminance retention and structural consistency, but still have limitations in detail recovery. Specifically, DWMFusion achieves an MI of 1.4016, and while it performs well in some areas, its Q_{MI} is only 0.4187, showing weaknesses in detail preservation and structural clarity compared to FDPMambaFuse. Mambadfuse and MMIF-INet exhibit similar performance. Mambadfuse achieves a Q_{MI} of 0.5862 and an $SSIM$ of 0.6608, showing some improvements in detail and texture retention, but still lacks cross-modal information complementarity and detail clarity. MMIF-INet in the MRI-SPECT task achieves a $VIFF$ of 0.8348, performing well in some areas, but its $SSIM$ is 0.5298 and its Q_{MI} is 0.5147, indicating clear limitations in detail preservation and perceptual quality. In conclusion, FDPMambaFuse not only preserves structural details and luminance information better in MRI-PET and MRI-SPECT fusion tasks but also surpasses existing SOTA methods in multiple key metrics, further demonstrating its superiority in MMIF.

4.3.2. HIT-MMIF-PAUS image fusion

Fig. 4 presents the qualitative fusion results of various methods on the HIT-MMIF-PAUS dataset, covering three imaging modalities: line-array, ring-array, and endoscopic imaging, with three fusion examples provided for each modality. In the overall fusion task on this dataset, FDPMambaFuse achieves the highest scores across multiple key metrics, including EN of 4.4448, MI of 1.0659, Q_{MI} of 0.3920, SD of 67.2982, AG of 7.3720, $SSIM$ of 0.5798, and $VIFF$ of 0.8245, demonstrating the best performance in structural detail preservation, luminance dynamics, and perceptual fidelity, significantly outperforming other methods. In the line-array fusion task, the first two cases represent subcutaneous tumor imaging, while the third corresponds to skin imaging. FDPMambaFuse achieves smooth structural transitions between PA and US modalities, reducing blurring and artifacts while clearly preserving key details in tumor trophoblast layers and vascular regions, enhancing the recognizability and consistency of pathological structures. In contrast to other methods, CSMCA, NSST-PAPCNN, IFCNN, U2Fusion, and SwinFusion, despite emphasizing PA features, exhibit significant deficiencies in preserving US luminance and structural information. MM-Net fails to adequately preserve PA details when additional US features are present. Although Mambadfuse improves texture representation, it still tends to introduce artifacts in vascular regions, compromising visual consistency. In the ring-array imaging task, all three cases involve imaging of the fingers. FDPMambaFuse balances the luminance of PA and US features while maintaining clear vascular and bone boundaries, significantly

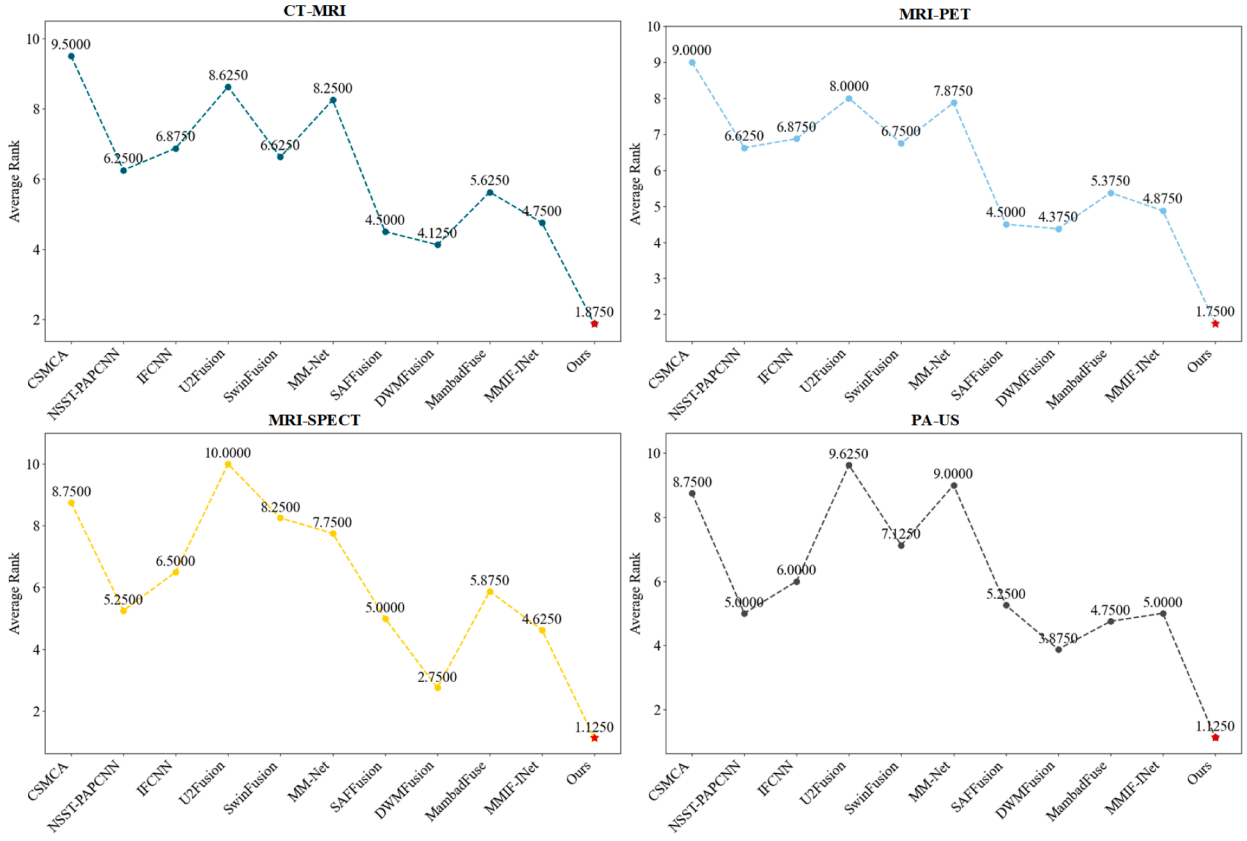

Fig. 5. Average ranking of different methods on four fusion modalities.

Table 2
Ablation study of DWT in deep feature extraction. The best score is marked in **bold**.

No.	Method	$EN\uparrow$	$MI\uparrow$	$Q_{MI}\uparrow$	$SD\uparrow$	$AG\uparrow$	$Q_{ABF}\uparrow$	$SSIM\uparrow$	$VIFF\uparrow$
E.0	–	5.48	1.34	0.42	89.62	14.25	0.48	0.51	0.58
E.1	w/ FFT	5.65	1.39	0.48	89.97	14.76	0.49	0.57	0.65
E.2	w/ DWT (Haar)	5.92	1.45	0.56	90.76	14.88	0.49	0.63	0.70
E.3	w/ DWT (Symlet-4)	6.04	1.47	0.55	91.14	14.89	0.51	0.64	0.71
E.4	w/ DWT (2 layers)	6.02	1.45	0.56	91.25	14.92	0.51	0.62	0.69
E.5	w/ DWT (3 layers)	5.95	1.44	0.57	91.28	14.90	0.51	0.65	0.72
E.6	Ours	**6.08**	**1.48**	**0.58**	**91.34**	**14.94**	**0.52**	**0.66**	**0.73**

Table 3
Ablation study of Mamba variants and multi-path scanning strategies in deep feature extraction. The best score is marked in **bold**.

No.	Method	$EN\uparrow$	$MI\uparrow$	$Q_{MI}\uparrow$	$SD\uparrow$	$AG\uparrow$	$Q_{ABF}\uparrow$	$SSIM\uparrow$	$VIFF\uparrow$	FLOPs(G)↓	Param.(M)↓
E.7	Convolution Block	5.82	1.28	0.42	87.92	12.25	0.35	0.51	0.54	**53.62**	**2.14**
E.8	Transformer	5.98	1.34	0.47	89.25	13.64	0.36	0.54	0.61	71.45	10.45
E.9	Mamba	6.02	1.39	0.49	90.44	14.72	0.36	0.55	0.62	57.89	8.55
E.10	Two-branch Mamba	6.03	1.41	0.49	90.67	14.77	0.40	0.57	0.65	57.89	4.36
E.11	Four-branch Mamba (route 1)	6.06	1.45	0.53	91.18	14.89	0.48	0.61	0.69	57.89	2.25
E.12	Four-branch Mamba (route 2)	6.04	1.43	0.54	91.14	14.83	0.46	0.60	0.69	57.89	2.25
E.13	Four-branch Mamba (route 3)	6.05	1.46	0.54	91.23	14.87	0.48	0.62	0.68	57.89	2.25
E.14	Four-branch Mamba (route 4)	6.06	1.46	0.57	91.28	14.91	0.51	0.64	0.70	57.89	2.25
E.6	MSD-Mamba	**6.08**	**1.48**	**0.58**	**91.34**	**14.94**	**0.52**	**0.66**	**0.73**	57.89	2.25

reducing noise and blurring. In comparison, SAFFusion, DWMFusion, and MMIF-Net often suffer from luminance loss, while Mambadfuse shows slight advantages in edge texture retention but introduces noticeable fusion noise. In the endoscopic imaging task, all three cases correspond to imaging of the rectum. FDPMambaFuse outperforms other methods in preserving tissue boundaries and mucosal textures, clearly delineating lesion areas, and enhancing clinical interpretability.

4.3.3. *Comparative fusion result summary*

FDPMambaFuse demonstrates consistently strong performance across all MMIF tasks. As shown in Figs. 3 and 4, the method achieves superior qualitative performance in both the brain-atlas fusion tasks and the HIT-MMIF-PAUS task. It effectively preserves structural details, enhances boundary clarity, and integrates complementary information from different modalities. These advantages are particularly evident in

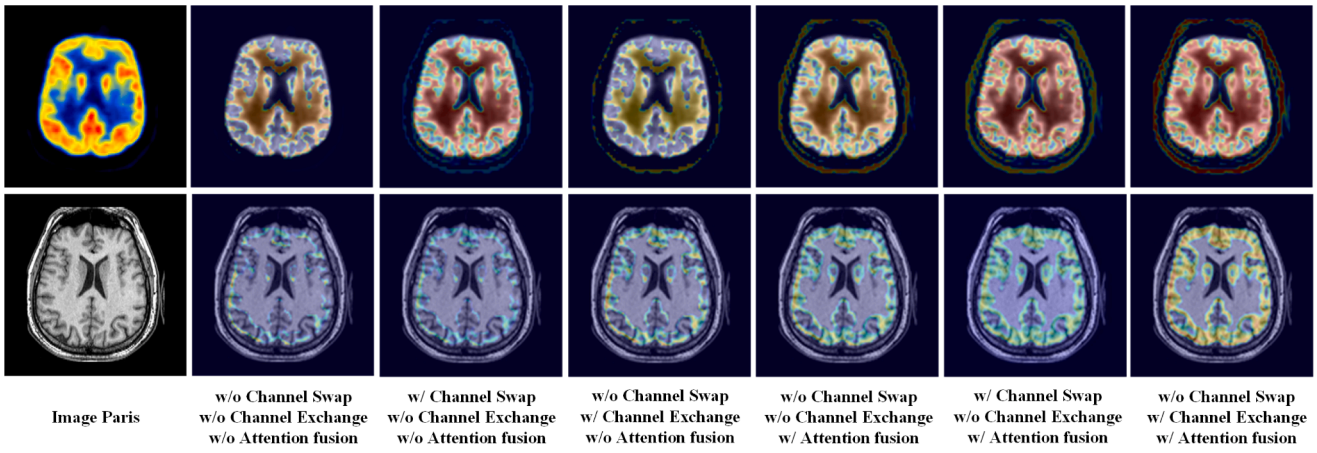

Fig. 6. Grad-CAM visualization of different strategies of the feature fusion module.

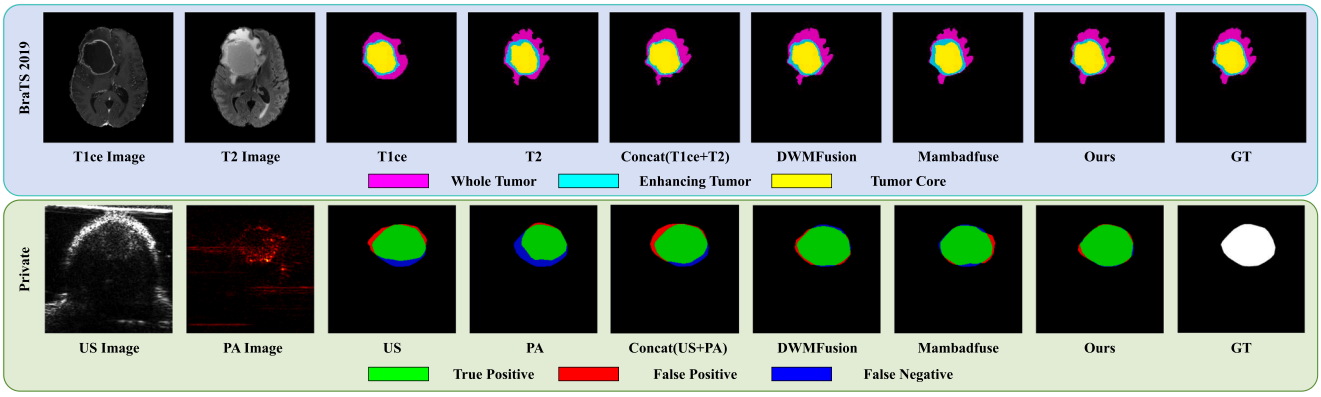

Fig. 7. Segmentation visualization results of different methods on the BraTS 2019 dataset and the private PA-US tumor segmentation dataset.

Table 4
Ablation study of fusion strategies. The best score is marked in **bold**.

No.	Channel Swap	Channel Exchange	Attention Mechanism	$EN\uparrow$	$MI\uparrow$	$Q_{MI}\uparrow$	$SD\uparrow$	$AG\uparrow$	$Q_{ABF}\uparrow$	$SSIM\uparrow$	$VIFF\uparrow$
E.15	✗	✗	✗	6.03	1.43	0.52	91.15	14.82	0.45	0.61	0.69
E.16	✗	✓	✗	6.04	1.44	0.53	91.22	14.85	0.47	0.64	0.70
E.17	✓	✗	✗	6.03	1.43	0.52	91.18	14.83	0.45	0.62	0.69
E.18	✗	✗	✓	6.07	1.46	0.56	91.26	14.88	0.49	0.65	0.72
E.19	✓	✗	✓	6.07	1.46	0.57	91.28	14.90	0.51	0.65	0.73
E.6	✗	✓	✓	**6.08**	**1.48**	**0.58**	**91.34**	**14.94**	**0.52**	**0.66**	**0.73**

Table 5
Ablation study of loss functions and uncertainty-guided loss components. The best score is marked in **bold**.

No.	Loss Function	$EN\uparrow$	$MI\uparrow$	$Q_{MI}\uparrow$	$SD\uparrow$	$AG\uparrow$	$Q_{ABF}\uparrow$	$SSIM\uparrow$	$VIFF\uparrow$
E.20	w/ $\mathcal{L}_{un}$ only	5.70	1.40	0.46	89.80	14.51	0.50	0.61	0.68
E.21	w/ $\mathcal{L}_{grad}$ only	5.65	1.36	0.43	89.63	14.65	0.51	0.65	0.65
E.6	w/ $(\mathcal{L}_{un} + \mathcal{L}_{grad})$	**6.08**	**1.48**	**0.58**	**91.34**	**14.94**	**0.52**	**0.66**	**0.73**
E.22	w/ pseudo-target only (no U-weight) + $\mathcal{L}_{grad}$	6.01	1.44	0.54	90.21	14.89	0.51	0.65	0.71
E.23	w/ U-weight only (fixed T) + $\mathcal{L}_{grad}$	6.04	1.45	0.52	90.92	14.91	0.50	0.65	0.71
E.24	w/ residual-weighted (post-fusion residual R) + $\mathcal{L}_{grad}$	5.82	1.41	0.48	89.82	14.87	0.51	0.64	0.69

the clearer and more coherent representation of clinically important structures, including bone, vasculature, and soft tissues. Quantitative results in Table 1 further confirm the superiority of FDPMambaFuse, as it achieves the highest scores on several key metrics, highlighting its strengths in information preservation, texture enhancement, structural fidelity, and perceptual quality. Moreover, the average ranking results in Fig. 5 show that FDPMambaFuse ranks first across all four tasks, demonstrating excellent stability and robustness. Overall, FDPMambaFuse delivers state-of-the-art fusion performance in both visual quality and quantitative evaluation, making it highly suitable for scenarios involving complex anatomical structures and large modality discrepancies. Its strong generalizability also suggests promising potential for broader clinical applications.

4.4. Ablation study

To evaluate the effectiveness of key components in the proposed fusion model, a series of ablation experiments are conducted to assess the

impact of different configurations on model performance. All ablation experiments were conducted on the MRI-PET dataset.

4.4.1. Analysis of deep feature extraction

We conduct ablation experiments to evaluate the contribution of the DWT-based Parallel Mamba module in the proposed deep feature extractor, as shown in Table 2. Compared to spatial-domain modeling methods (E.0), FFT- and DWT-based frequency-domain feature extraction methods improve all metrics, particularly information retention and structural fidelity. However, DWT outperforms FFT in fusion performance by effectively capturing multi-scale structural and textural information. Further experiments reveal that the choice of wavelet basis significantly affects fusion performance. Using the Daubechies-4 wavelet achieves the best results across all metrics, particularly in Q_{MI}, $SSIM$, and $VIFF$, demonstrating superior structural consistency and detail retention. Additionally, using multi-layer DWT does not improve fusion performance compared to a single-layer approach and instead increases computational cost. Therefore, the proposed method, which uses the single-layer DWT with the Daubechies-4 wavelet, achieves the best fusion results.

Secondly, to further validate the effectiveness of the proposed MSD-Mamba, we perform a systematic comparison among the convolution block, Transformer, standard Mamba, two-branch parallel Mamba, four-branch parallel Mamba with the single scanning direction, and MSD-Mamba. The experimental results are summarized in Table 3. Compared with the convolution block (E.7), both the Transformer (E.8) and Mamba (E.9) achieve substantial improvements across most fusion metrics, indicating that architectures with long-range dependency modeling better capture cross-modal correlations and enhance structural consistency and detail fidelity. However, the Transformer incurs substantially higher FLOPs and parameter counts, leading to reduced computational efficiency. In contrast, Mamba achieves superior fusion performance while maintaining a lightweight structure. Building on this, the two-branch parallel Mamba (E.10) yields further improvements across most metrics, demonstrating that multi-branch state modeling enhances feature representation. When the architecture is extended to the four-branch parallel form (E.11-E.14), fusion performance continues to improve, and the parameter count decreases to 2.25M, representing a 73.68% reduction compared with the single-branch Mamba (E.9), indicating higher structural efficiency and more effective information utilization. Moreover, the four scanning paths in the four-branch parallel design show slight differences in fusion results, indicating that the scanning direction influences feature modeling and, consequently, affects the final fusion performance. Finally, the proposed MSD-Mamba (E.6), which integrates multiple scanning-path strategies, achieves the best performance across all core metrics, confirming the effectiveness of its architectural design.

4.4.2. Analysis of multi-stage feature fusion

To evaluate the structural roles of different components in the proposed multi-stage fusion framework, we conduct ablation experiments, as summarized in Table 4. The baseline model (E.15), without any fusion mechanism, shows inferior performance across most metrics, indicating limited cross-modal interaction under simple aggregation. Introducing channel exchange alone (E.16) consistently improves structural and perceptual metrics, demonstrating that cross-modal topology reorganization enhances modality interaction even without adaptive modulation. Channel swap (E.17) provides marginal gains, suggesting that simple channel-level reallocation is less effective than explicit topology restructuring. Applying attention alone (E.18) also improves performance, confirming the benefit of spatially adaptive modulation. However, its performance remains below the complete model (E.6), indicating that adaptive modulation benefits from prior topology reorganization for more effective fusion. When combining channel swap with attention (E.19), further improvements are observed, yet the results are still slightly inferior to the full configuration. The complete strategy (E.6), integrating topology reorganization and adaptive modulation, achieves the best performance across all metrics, particularly in structural consistency (SD), mutual information (MI), and perceptual quality indicators (Q_{MI}, $VIFF$). These results indicate that effective fusion arises from the coordinated interaction between topology restructuring and spatial modulation, highlighting the complementary roles of structural modeling and adaptive modulation. Grad-CAM visualizations (Fig. 6) further confirm that the complete model produces more coherent activation patterns in anatomically significant regions.

Table 6
Comparison of different methods on the BraTS 2019 and private PA-US tumor segmentation datasets. For BraTS 2019, WT, ET, TC, and mean Dice (%) are reported, while only mean Dice (%) is reported for the PA-US dataset. The best score is marked in **bold**.

Method	BraTS 2019				Private
	WT (%)↑	ET (%)↑	TC (%)↑	mDice (%)↑	mDice (%)↑
Source Image A	65.72	81.32	74.87	73.97	70.61
Source Image B	83.25	55.43	69.45	69.38	73.25
Concat(A + B)	81.72	83.13	82.26	82.37	78.33
DWMFusion	84.95	83.96	82.41	83.77	80.92
Mambadfuse	84.88	83.62	82.34	83.61	81.21
Ours	**85.76**	**84.15**	**83.26**	**84.39**	**82.36**

4.4.3. Analysis of loss function

To validate the effectiveness of the proposed uncertainty-guided hybrid loss, we conduct systematic ablation experiments, and the results are summarized in Table 5.

We first compare the basic configurations, including using only the uncertainty loss ($\mathcal{L}_{un}$), only the gradient loss ($\mathcal{L}_{grad}$), and their combination. The results show that $\mathcal{L}_{un}$ outperforms $\mathcal{L}_{grad}$ in information retention and structural consistency, achieving higher Q_{MI} and SD values. In contrast, the gradient loss alone mainly focuses on edge consistency and is less effective in maintaining cross-modal semantic agreement. The complete hybrid loss ($\mathcal{L}_{un} + \mathcal{L}_{grad}$) achieves the best overall performance across all evaluation metrics, particularly in fusion quality and detail preservation, confirming the complementary effects of the two loss components.

To further analyze the internal mechanism of the uncertainty-guided loss, we decompose it into two functional components. In E.22, the pseudo-supervision target is retained while removing the uncertainty-based weighting in the loss, using a standard MSE formulation instead. In E.23, the uncertainty weighting is preserved, but the pseudo-supervision target is replaced with a fixed average fusion result ($T = 0.5I_A + 0.5I_B$). In E.24, a conventional residual-weighted baseline is implemented, where post-fusion residuals between the fused image and the input modalities are used as weighting factors. The results indicate that both E.22 and E.23 lead to consistent performance degradation compared with the complete formulation, demonstrating that pseudo-supervision construction and confidence-weighted optimization jointly contribute to performance improvement. Moreover, the residual-weighted baseline (E.24) further underperforms the proposed formulation, indicating that the performance gain stems from the coordinated design of supervision construction and optimization regulation.

4.5. Downstream applications

To further validate the effectiveness of FDPMambaFuse in improving downstream task performance, we conduct comprehensive evaluations on both the BraTS 2019 multi-parametric MRI brain tumor segmentation dataset and our private PA-US tumor segmentation dataset, assessing how the fused images contribute to segmentation accuracy. In BraTS 2019, the T1ce and T2 MRI modalities are selected as fusion inputs, whereas in the private dataset, PA and US modalities are used. All experiments employ nnU-Net (Isensee et al., 2018) as the downstream segmentation network, and comparisons are conducted using the source images and the fused images generated by the top three fusion methods

based on average performance. All training configurations are kept identical to ensure fairness and reproducibility, and the Dice coefficient is used as the evaluation metric. Quantitative results are reported in Table 6, and qualitative visualization results are shown in Fig. 7.

In the BraTS 2019 dataset, T1ce provides stronger responses for enhancing tumor (ET) and tumor core (TC), whereas T2 is more sensitive to the whole tumor region (WT). After multimodal fusion, FDPMambaFuse achieves superior segmentation performance for ET, TC, and WT compared with DWMFusion and MambaDFuse. Qualitative results further show that FDPMambaFuse produces clearer and more accurate delineations of tumor boundaries. In the private PA-US dataset, PA highlights high-absorption features such as vasculature, whereas US provides clearer visualization of soft tissue structures and tumor margins. The fused images combine these complementary advantages, and FDPMambaFuse consistently outperforms other methods across all segmentation metrics, demonstrating the strongest enhancement effect on the downstream task. Overall, FDPMambaFuse significantly improves tumor segmentation quality on both datasets, validating the practical value of its fusion results in real clinical scenarios.

5. Discussion

Despite FDPMambaFuse performing excellently in most MMIF tasks, we found that PA and US modalities are still affected by noise interference during fusion. These noise sources, including low-SNR regions, detector noise, and tissue scattering artifacts, limit the model's performance in low-SNR regions and affect the extraction of key information. Therefore, future work will explore the introduction of an image filtering enhancement module in FDPMambaFuse to adaptively improve the quality of low-SNR regions, reduce noise interference, and further enhance the structural integrity and visual quality of the fused images. Additionally, although FDPMambaFuse performs excellently with well-aligned multi-modal images, it still has limitations in handling misaligned images or scale differences. In clinical practice, image misalignment or anatomical region differences can negatively affect the fusion results. Therefore, future research will focus on incorporating image alignment mechanisms to enhance the model's adaptability in non-rigid or weak alignment conditions, expanding its applicability in complex clinical environments.

6. Conclusion

In this paper, we present FDPMambaFuse, a novel MMIF framework that integrates frequency-domain decomposition with multi-directional parallel Mamba to achieve efficient and precise fusion. The framework employs a multi-level CNN-Mamba feature extractor to enable effective spatial-frequency-directional representation, and introduces a multi-stage fusion strategy to enhance cross-modal semantic complementarity and feature aggregation. Meanwhile, we design an unsupervised hybrid loss that combines pixel-level uncertainty with gradient-based structural constraints to improve structural consistency and visual quality. In addition, we construct and publicly release HIT-MMIF-PAUS, a clinically meaningful PA-US fusion dataset. Extensive experiments on multiple MMIF benchmarks demonstrate that FDPMambaFuse outperforms state-of-the-art methods in fusion quality, structural fidelity, and edge clarity, and further achieves strong performance in downstream tumor segmentation, confirming its potential for clinical applications. Future work will investigate adaptive denoising and cross-modal alignment strategies to further improve robustness under low-SNR and weakly aligned conditions.

Declaration of generative AI in scientific writing

While preparing this work, the author(s) utilized DeepL and ChatGPT to enhance readability and language. Following the use of this tool/service, the author(s) reviewed and revised the content as required, and assume(s) complete responsibility for the publication's content.

CRediT authorship contribution statement

Baiya Li: Writing – original draft, Validation, Software, Methodology, Conceptualization; **Boheng Zhang:** Writing – original draft, Software, Methodology, Data curation; **Yang Liu:** Visualization, Formal analysis; **Haorui Huang:** Resources, Data curation; **Cailing Lin:** Investigation; **Rui Fan:** Writing – review & editing, Supervision; **Mingjian Sun:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by the *National Key R&D Program of China* [Grant No. 2023YFC2412100], the *National Natural Science Foundation of China* [Grant No. 62275062], the *Project of Shandong Innovation and Startup Community of High-end Medical Apparatus and Instruments* [Grant Nos. grantnumGS5011000018092023-SGTTXM-002 and 2024-SGTTXM-005], the *Shandong Province Technology Innovation Guidance Plan (Central Leading Local Science and Technology Development Fund)* [Grant No. YDZX2023115], the *Taishan Scholar Special Funding Project of Shandong Province*, and the *Shandong Laboratory of Advanced Biomaterials and Medical Devices in Weihai* [Grant No. ZL202402].

References

Azam, M. A., Khan, K. B., Salahuddin, S., Rehman, E., Khan, S. A., Khan, M. A., Kadry, S., & Gandomi, A. H. (2022). A review on multimodal medical image fusion: Compendious analysis of medical modalities, multimodal databases, fusion techniques and quality metrics. *Computers in Biology and Medicine*, *144*, 105253.

Bashir, R., Junejo, R., Qadri, N. N., Fleury, M., & Qadri, M. Y. (2019). SWT and PCA image fusion methods for multi-modal imagery. *Multimedia Tools and Applications*, *78*, 1235–1263.

Di, J., Guo, W., Liu, J., Ren, L., & Lian, J. (2024). AmmNet: A multimodal medical image fusion method based on an attention mechanism and MobileNetV3. *Biomedical Signal Processing and Control*, *96*, 106561. https://doi.org/10.1016/j.bspc.2024.106561

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S. et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fan, C., Yun, M., Chen, L., Zhao, H., & Peng, B. (2025). DWMFusion: A lightweight medical image fusion network based on discrete wavelet convolution and selective state space models. *Expert Systems with Applications*, *286*, 128040. https://doi.org/10.1016/j.eswa.2025.128040

Fang, S., Li, K., & Li, Z. (2023). Changer: Feature interaction is what you need for change detection. *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–11.

Fu, J., Li, W., Du, J., & Huang, Y. (2021). A multiscale residual pyramid attention network for medical image fusion. *Biomedical Signal Processing and Control*, *66*, 102488.

Grosu, A. L., Weber, W. A., Franz, M., Stärk, S., Piert, M., Thamm, R., Gumprecht, H., Schwaiger, M., Molls, M., & Nieder, C. (2005). Reirradiation of recurrent high-grade gliomas using amino acid PET (SPECT)/CT/MRI image fusion to determine gross tumor volume for stereotactic fractionated radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, *63*(2), 511–519. https://doi.org/10.1016/j.ijrobp.2005.01.056

Guihong, Q., Dali, Z., & Pingfan, Y. (2001). Medical image fusion by wavelet transform modulus maxima. *Optics Express*, *9*(4), 184.

He, D., Li, W., Wang, G., Huang, Y., & Liu, S. (2025). MMIF-INet: Multimodal medical image fusion by invertible network. *Information Fusion*, *114*, 102666.

Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A review of multimodal medical image fusion techniques. *Computational and Mathematical Methods in Medicine*, *2020*(8), 1–16.

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P. F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S. et al. (2018). nnU-Net: Self-adapting framework for U-net-based medical image segmentation. arXiv preprint arXiv:1809.10486.

Jimenez-Mesa, C., Arco, J. E., Martinez-Murcia, F. J., Suckling, J., Ramirez, J., & Gorriz, J. M. (2023). Applications of machine learning and deep learning in SPECT and PET imaging: General overview, challenges and future prospects. *Pharmacological Research*, *197*, 106984. https://doi.org/10.1016/j.phrs.2023.106984

Kim, J., Heo, D., Cho, S., Ha, M., Park, J., Ahn, J., Kim, M., Kim, D., Jung, D. H., & Kim, H. H. (2024). Enhanced dual-mode imaging: Superior photoacoustic and ultrasound endoscopy in live pigs using a transparent ultrasound transducer. *Science Advances*, *10*(47), eadq9960.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y. et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4015–4026).

Li, G., Huang, Q., Wang, W., & Liu, L. (2025). Selective and multi-scale fusion mamba for medical image segmentation. *Expert Systems with Applications*, *261*, 125518.

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

Li, X., Zhou, F., Tan, H., Zhang, W., & Zhao, C. (2021). Multimodal medical image fusion based on joint bilateral filter and local gradient energy. *Information Sciences*, *569*, 302–325.

Li, Y., Zhang, Y., Timofte, R., Van Gool, L., Tu, Z., Du, K., Wang, H., Chen, H., Li, W., Wang, X. et al. (2023). Ntire 2023 Challenge on image denoising: Methods and results. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 1905–1921).

Li, Z., Pan, H., Zhang, K., Wang, Y., & Yu, F. (2024). MambaDFuse: A Mamba-based dual-phase model for multi-modality image fusion. arXiv preprint arXiv:2404.08406.

Lin, L., Xie, J., Wang, Y., Huang, J., Zhuang, R., Tu, X., Ding, X., Shen, N., & Lu, Q. (2025). Spatial-frequency dual-domain kolmogorov–arnold networks for multimodal medical image fusion. *Neurocomputing*, *648*, 130661. https://doi.org/10.1016/j.neucom.2025.130661

Liu, J., Dian, R., Li, S., & Liu, H. (2023a). SGFusion: A saliency guided deep-learning framework for pixel-level image fusion. *Information Fusion*, *91*, 205–214.

Liu, R., Liu, Y., Wang, H., Li, J., & Hu, K. (2025). SAFFusion: A saliency-aware frequency fusion network for multimodal medical image fusion. *Biomedical Optics Express*, *16*(6), 2459–2481. https://doi.org/10.1364/BOE.555458

Liu, Y., Chen, X., Ward, R. K., & Wang, Z. J. (2019). Medical image fusion via convolutional sparsity based morphological component analysis. *IEEE Signal Processing Letters*, *26*(3), 485–489.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Jiao, J., & Liu, Y. (2024). VMamba: Visual state space model. *Advances in Neural Information Processing Systems*, *37*, 103031–103063.

Liu, Y. et al. (2023b). An improved hybrid network with a transformer module for medical image fusion. *IEEE Journal of Biomedical and Health Informatics*, *27*(7), 3489–3500. https://doi.org/10.1109/JBHI.2023.3264819

LiuY., ChenY., ChengJ., Jane, W., & ChenX. (2024). MM-Net: A mixformer-based multi-scale network for anatomical and functional image fusion. *IEEE Transactions on Image Processing*, *33*, 2197–2212.

Luo, F., Wu, D., Pino, L. R., & Ding, W. (2025). A novel multimodel medical image fusion framework with edge enhancement and cross-scale transformer. *Scientific Reports*, *15*, 11657.

Ma, J., Li, F., & Wang, B. (2024). U-Mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv preprint arXiv:2401.04722.

Ma, J., Tang, L., Fan, F., Huang, J., Mei, X., & Ma, Y. (2022). SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer. *IEEE/CAA Journal of Automatica Sinica*, *9*(7), 18.

Park, J., Choi, S., Knieling, F., Clingman, B., Bohndiek, S., Wang, L. V., & Kim, C. (2024). Clinical translation of photoacoustic imaging. *Nature Reviews Bioengineering*, *3*, 1–20.

Parmar, K., & Kher, R. (2012). A comparative analysis of multimodality medical image fusion methods. In *Modelling symposium (AMS), 2012 Sixth asia*.

Rao, W., Gao, L., Qu, Y., Sun, X., Zhang, B., & Chanussot, J. (2022). Siamese transformer network for hyperspectral image target detection. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–19.

Satyanarayana, V., & Mohanaiah, P. (2025). Multimodal medical image fusion in NSST domain in coupled neural systems. *International Journal of Computational Intelligence Systems*, *18*, 111. https://doi.org/10.1007/s44196-025-00841-4

Singh, R., Vatsa, M., & Noore, A. (2009). Multimodal medical image fusion using redundant discrete wavelet transform. In *2009 Seventh international conference on advances in pattern recognition* (pp. 232–235). IEEE.

Srivastava, S., Bhatia, S., Agrawal, A. P. et al. (2025). Deep adaptive fusion with cross-modality feature transition and modality quaternion learning for medical image fusion. *Evolving Systems*, *16*, 17. https://doi.org/10.1007/s12530-024-09648-8

Stokking, R., Zuiderveld, K. J., & Viergever, M. A. (2001). Integrated volume visualization of functional image data and anatomical surfaces using normal fusion. *Human Brain Mapping*, *12*(4), 203–218.

Tang, W., He, F., Liu, Y., & Duan, Y. (2022a). Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, *31*, 5134–5149. https://doi.org/10.1109/TIP.2022.3193288

Tang, W., He, F., Liu, Y., & Duan, Y. (2022b). Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, *31*, 5134–5149.

Wang, C., Nie, R., Cao, J., Wang, X., & Zhang, Y. (2022). IGNFusion: An unsupervised information gate network for multimodal medical image fusion. *IEEE Journal of Selected Topics in Signal Processing*, *16*(4), 16.

Wen, J., Qin, F., Du, J., Fang, M., Wei, X., Chen, C. L. P., & Li, P. (2023). MSGFusion: Medical semantic guided two-branch network for multimodal brain image fusion. *IEEE Transactions on Multimedia*, *26*, 944–957.

Xu, H., Ma, J., Jiang, J., Guo, X., & Ling, H. (2020). U2Fusion: A unified unsupervised image fusion network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 502–518.

Yang, B., & Li, S. (2009). Multifocus image fusion and restoration with sparse representation. *IEEE Transactions on Instrumentation and Measurement*, *59*(4), 884–892.

Yang, Z., & Farsiu, S. (2023). Directional connectivity-based segmentation of medical images. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition* (pp. 11525–11535).

Ye, S., Wang, T., Ding, M., & Zhang, X. (2023). F-DARTS: Foveated differentiable architecture search based multimodal medical image fusion. *IEEE Transactions on Medical Imaging*, *42*, 3348–3361.

Yin, M., Liu, X., Liu, Y., & Chen, X. (2018). Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain. *IEEE Transactions on Instrumentation and Measurement*, *68*(1), 49–64.

Zhang, B., Huang, H., Shen, Y., & Sun, M. (2025). MM-UKAN + +: A novel Kolmogorov-Arnold network based u-shaped network for ultrasound image segmentation. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, *72*, 498–514.

Zhang, Q., Liu, Y., Blum, R. S., Han, J., & Tao, D. (2018). Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, *40*, 57–75.

Zhang, Y., Liu, Y., Sun, P., Yan, H., Zhao, X., & Zhang, L. (2020). IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, *54*, 99–118.

Zhou, M. et al. (2025). A general spatial-frequency learning framework for multimodal image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(7), 5281–5298. https://doi.org/10.1109/TPAMI.2024.3368112

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., & Wang, X. (2024). Vision Mamba: Efficient visual representation learning with bidirectional state space model. https://arxiv.org/abs/2401.09417.