Biomedical Signal Processing and Control 103 (2025) 107357

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

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

AFPNet: An adaptive frequency-domain optimized progressive medical image fusion network

Dangguo Shao, Hongjuan Yang, Lei Ma *, Sanli Yi

Faculty of Information Engineering and Automation, KunmingUniversity of Science and Technology,Kunming 650500, Yunnan, China

ARTICLE INFO

Keywords:
Medical Image Fusion
Adaptive Frequency-Domain Optimization
Progressive Feature Fusion Strategy
Attention Convolutional Networks

ABSTRACT

This study presents AFPNet, a novel Progressive Medical Image Fusion Network incorporating Adaptive Frequency-Domain Optimization, designed to enhance the fusion process of multi-modal medical imaging modalities, including SPECT-MRI and PET-MRI. AFPNet exploits a progressive Attention Convolutional Neural Network(ACNN) to significantly enhance the quality of fused medical images. Key innovations encompass a dual-branch module for efficient spatial and channel feature fusion, a Global Enhancement Attention Module for the seamless integration of global information, and a multi-scale feature fusion strategy to capture diverse contextual information. Experimental results on the test datasets indicate that AFPNet generally outperforms state-of-the-art approaches in key metrics such as Mutual Information (MI) and Visual Information Fidelity (VIF), showing notable improvements in preserving both structural coherence and fine-grained details. The incorporation of adaptive frequency-based weighting mechanisms within the model optimize fusion of high-frequency and low-frequency components, rendering it well-suited for medical imaging applications demanding exceptional precision.

1. Introduction

As medical imaging technology continues to advance, medical images are becoming increasingly integral to clinical diagnosis. However, owing to the inherent differences in imaging principles, the information conveyed by medical images varies significantly. Magnetic Resonance Imaging (MRI) offers high-resolution anatomical details, meticulously rendering structures such as organs, tissues, and blood vessels, thereby facilitating the observation of anatomical changes and organ morphology. Conversely, Positron Emission Tomography (PET) and Single Photon Emission Computed Tomography (SPECT) furnish functional insights into metabolic processes, blood flow distribution, and other physiological parameters. To address the inherent limitations of single-modality medical images, the technique of medical image fusion [1,2] has been developed. By synergistically integrating complementary information from different imaging modalities, medical image fusion produces richer composite images, offering a more comprehensive foundation for clinical diagnosis and enabling physicians to more accurately localize lesions and formulate treatment plans. For instance, SPECT-MRI image fusion amalgamates metabolic functional data with anatomical structural details, presenting substantial advantages in the diagnosis of neurological disorders, cardiovascular conditions, and tumors. Similarly, MRI-PET image fusion enhances image contrast, thereby facilitating the detection of abnormal regions and lesions, which is essential for the diagnosis of tumors and neurological disorders.

Current image fusion methodologies can be broadly categorized into traditional techniques and deep learning-based approaches. Traditional techniques encompass spatial domain methods, such as weighted averaging [3] and Principal Component Analysis (PCA) [4], which excel in processing low-frequency information but often falter when handling high-frequency components, such as edges and details. Additionally, frequency domain approaches, such as Wavelet Transform (WT) [5], Curved Wave Transform (CVT) [6] are generally more proficient at preserving image details and structural integrity. Sparse representation-based methods [7] leverage sparsity theory, representing images as sparse linear combinations of dictionary elements for fusion. However, these methods predominantly depend on manually engineered feature extraction techniques and fusion rules, which, despite their effectiveness in certain scenarios, tend to become increasingly complex as the demand for superior fusion performance escalates, consequently resulting in greater computational resource consumption and diminished efficiency.

In contrast to traditional methods, deep learning-based medical

* Corresponding author.
E-mail address: 1974997005@qq.com (L. Ma).

https://doi.org/10.1016/j.bspc.2024.107357
Received 18 September 2024; Received in revised form 1 December 2024; Accepted 16 December 2024
Available online 28 December 2024

image fusion approaches develop fusion models with robust generalization capabilities from extensive datasets and overcome the limitations associated with manual feature selection in conventional methods, thereby rendering the fusion process more automated and resilient. However, these approaches still face two major challenges when processing multimodal images. First, although Convolutional Neural Networks (CNNs) being excel in local feature extraction, their limited receptive field restrict effective cross-modal global information capture [8], which is crucial for effectively handling multi-modal medical images. Transformer-based models are capable of capturing global contextual information through self-attention mechanisms, enabling global dependency modeling. However, their high computational cost restricts their widespread applicability in practical scenarios. To address these limitation, some CNN-based fusion methods employ attention mechanisms to enhance global perception [9]. However, most methods directly concatenate source image before feature extraction [10] or extract features separately prior to fusion [11]. The former approach often leads to simple feature aggregation, hindering the full exploitation of the complementary characteristics of source images, thus affecting the structural integrity and detail representation of the fused images. The latter, while capturing unique information from each image, lacks inter-image interaction, making it challenging to preserve overall consistency throughout the fusion process.

To address these challenges, this paper proposes a progressive Attention Convolutional Neural Network (ACNN) for medical image fusion, which integrates channel, spatial, and multi-scale information layer by layer, aiming to comprehensively integrate the complementary information from different modalities, thus generating high-quality fused images., In the first layer, we design a Dual-Branch Depthwise Separable Attention Module (DB-DSCAM), which combines a lightweight attention mechanism with depthwise separable convolution, to extracts spatial features from single-modality images and performs channel feature fusion for multi-modality images. This module enables local cross-modal complementary information while extracting unique spatial features from each modality. For instance, Fig. 1**(a)** illustrates the spatial feature extraction results for MRI and PET images, while Figure Fig. 1**(b)** depicts the channel fusion outcomes for MRI-PET and SPECT-MRI. In the second layer, the Global Enhancement Attention Module (GEAM), improved based on the gradient channel attention block [12], further enhances cross-modal information fusion by integrating the fusion results from the provious layer with the global spatial information of the source images via generating a global weight map. This significantly improves the overall consistency of the fused images and the correlation of multimodal features. As illustrated in Fig. 2, this module's feature extraction effect significantly highlights the integration of global information, as compared to Fig. 1**(a)**. In the third layer, we design a multi-scale feature fusion module (MSFM) based on Parallel Dilated Convolutional [13], which further merges multi-scale contextual information from previous two layers and the source images, achieving efficient fusion of global structures and local details at a more advanced level. Additionlly, within this progressive feature fusion framework, the source images are input at each layer to capture unique modality information, enhancing the model's robustness under input variations and ensuring both the retention of source image characteristics and the effective fusion of detailed information.

Moreover, to further enhance model performance, we propose an adaptive frequency-domain optimization strategy that, in combination with a dynamic weighting mechanism, is used to define an adaptive frequency-feature weighted loss function. By applying Fast Fourier Transform(FFT) to convert the image from the spatial domain to the frequency domain and measuring the frequency characteristics of the input images,the model adaptively adjusts the weights of the input images to dynamically optimize the spatial-domain fusion of PET/SPECT functional information(e.g., metabolic activity, blood flow) and MRI anatomical details (e.g., tissue boundaries, anatomical structures) [15–18]. Unlike traditional spatial-domain methods, the frequency-domain weighting mechanism enhances fusion of high-frequency (e.g., edges, textures) and low-frequency (e.g., overall contrast) components

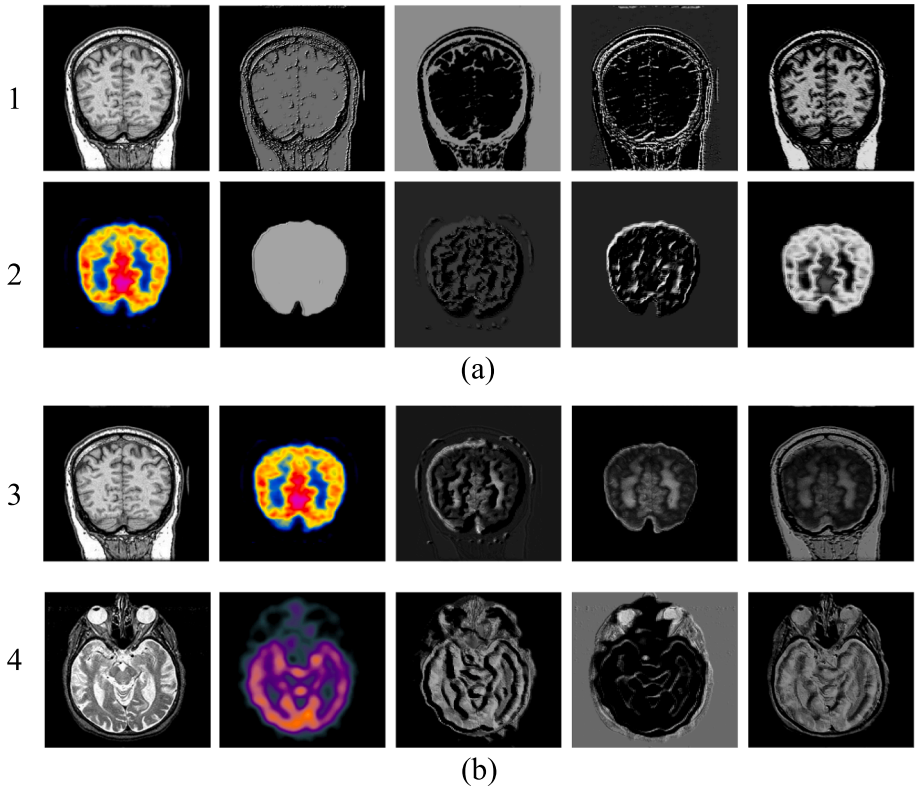


Fig. 1. Spatial feature extraction and channel fusion results based on lightweight attention mechanism and depthwise separable convolution.

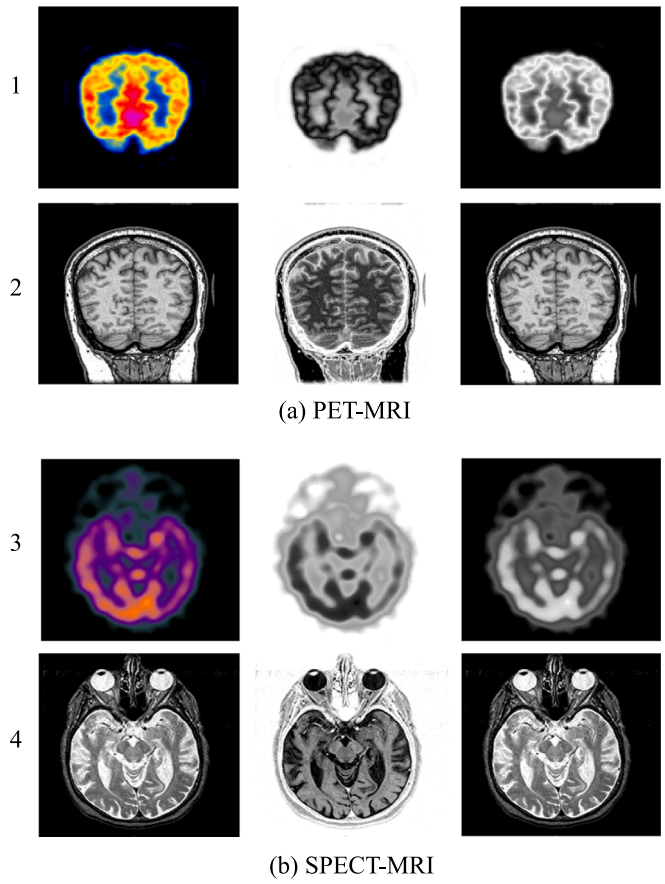


Fig. 2. The feature map extracted by the Global Enhanced Attention Module (GEAM).

from the frequency perspective. Research indicates that frequency-domain analysis methods, by analyzing an image's frequency components, can unveil subtle features that may be less detectable in the spatial domain and minimizing frequency-domain gaps can substantially enhance image reconstruction and synthesis quality [14].

Through a progressive, multi-stage and multi-level fusion mechanism, the model transitions from single-modality feature extraction to multimodal fusion. Combined with the joint optimization of the adaptive frequency-domain weighted loss function, the Progressive Medical Image Fusion Network with Adaptive Frequency-Domain Optimization (AFPNet) ensures that the final fused image maintains advantages in both detail preservation and consistency, significantly enhancing the overall image quality. The main contributions of this paper are summarized as follows:

1) **Module Design:** we design a Dual-Branch Depthwise Separable Attention Module (DB-DSCAM) to independently process channel and spatial information. This module enables spatial feature extraction for single-modal images and channel feature fusion for multimodal images. Additionally,the improved Global Enhanced Attention Module (GEAM) was designed to integrate cross-modal global spatial information. We also design a Multi-Scale Feature Fusion Module (MSFM) to merge multi-scale contextual information.
2) **Progressive Attention Fusion Framework:** we design a three-stage progressive attention fusion network by combining DB-DSCAM, GEAM, and MSFM with feature reuse strategies. This framework achieves efficient integration of channel, spatial, and multi-scale information through layer-by-layer feature fusion.
3) **Adaptive Frequency-Domain Optimization Loss:** we proposed an adaptive frequency-domain optimization strategy and further designed an adaptive frequency-domain weighted loss function by incorporating a dynamic weighting mechanism. By employing Fourier Transform to measure the frequency characteristics of input images, the model dynamically adjusts the loss function's weights, thereby optimizing the spatial-domain network and enhancing fusion performance.
4) **Application in Medical Image Fusion:** We validate the effectiveness of the model in PET-MRI and SPECT-MRI image fusion tasks, demonstrating its robust fusion capabilities.

2. Related work

This section provides a comprehensive review of prominent methods in medical image fusion (MIF). Specifically, Section 2.1 delineates traditional MIF approaches, Section 2.2 explores MIF techniques grounded in convolutional neural networks (ConvNet), and Section 2.3 examines MIF methodologies utilizing attention mechanisms.

2.1. Traditional MIF methods

Traditional medical image fusion methods can be broadly categorized into four primary types: spatial domain-based algorithms [19], transform domain-based algorithms, sparse representation-based methods [20], and hybrid methods.

Spatial domain-based algorithms manipulate image pixel values directly, with techniques such as weighted averaging [3] and principal component analysis (PCA) [4,21]. These methods perform fusion through linear or non-linear combinations of pixel values. For example, the weighted averaging method integrates pixel values from multi-modal images, while PCA extracts principal components to facilitate fusion. Zhang et al. [21] proposed an automated PCA approach that mitigates redundancy and amplifies significant information in the fused image. Transform domain-based algorithms generally utilize multiscale transform (MST) theory [22], incorporating common transforms such as discrete cosine transform (DCT) [23], Laplacian pyramid (LP) [24], wavelet transform (WT) [25], curvelet transform (CVT) [6], and non-subsampled contourlet transform (NSCT) [26]. These approaches decompose images into the frequency domain, apply fusion rules to the decomposed coefficients, and subsequently reconstruct the fused image through inverse transformation. Sparse representation-based methods leverage sparsity theory, representing images as sparse linear combinations of dictionary elements for fusion. Nejati et al. [27] introduced a method based on dictionary learning, wherein a redundant dictionary is constructed to capture the sparse characteristics of the image. The image is decomposed into sparse coefficients, which are subsequently fused. Hybrid methods amalgamate the strengths of diverse techniques to enhance fusion outcomes. For instance, Wang et al. [28] proposed a hybrid approach that integrates the benefits of the Laplacian pyramid and sparse representation, employing sparse representation to select features in high-frequency regions and the Laplacian pyramid to fuse global structures in low-frequency regions.

Nonetheless, traditional MIF algorithms depend on manually crafted decomposition methods and fusion strategies, which restrict their capacity to fully exploit complementary information within images. Consequently, this often leads to limited fusion performance and necessitates complex, time-intensive computations.

2.2. ConvNet-Based MIF algorithms

To address the limitations inherent in traditional methods, ConvNet-based MIF algorithms autonomously learn image features, thereby circumventing the need for intricate transformation selection and fusion strategy design. Principal approaches encompass sequential convolution methods [29,30], encoder-decoder (EN-DN) frameworks [31], and generative adversarial network (GAN)-based [32] techniques.

Liu et al. [33] pioneered the application of CNNs to image fusion,

proposing a sequential convolution-based multi-focus image fusion methodology. This approach autonomously extracts features from source images for fusion, with the fused image reconstructed via deconvolution. This methodology effectively obviates the need for manually crafted feature extraction and fusion strategies, thereby rendering image fusion more automated and robust. In encoder-decoder-based approaches, a pre-trained encoder extracts features from source images, which are then fused according to established rules, with the fused features reconstructed into a composite image by the decoder. For instance, SEDRFuse [34] utilizes a symmetric encoder-decoder architecture, preserving image details while leveraging residual blocks to augment edge and detail information in the fused images. GAN-based MIF algorithms produce high-quality fusion images through adversarial training involving generators and discriminators. Zhou et al. proposed DDcGAN [35], which employs dual discriminators to handle varying resolution levels of image layers, ensuring that the fused images preserve low-frequency structures while accentuating high-frequency details. MGMDcGAN [36] further advances fusion quality by extracting image features across multiple scales through the use of multiple generators and discriminators.

In contrast to traditional methods, ConvNet-based MIF approaches develop fusion models from extensive datasets, thereby significantly enhancing fusion performance. Nonetheless, due to the restricted receptive field of CNNs, the extraction of global complementary information remains suboptimal. This constraint may impede the effective utilization of global features in complex multi-modal medical image fusion tasks.

2.3. Attention Mechanism-Based MIF methods

Attention mechanism-based methods primarily encompass attention convolutional networks (ACNN) [37–39] and Transformer models [40–42]. These approaches leverage attention mechanisms to enhance feature extraction, particularly in addressing the limitations of traditional CNNs and capturing global information.

The ACNN integrate CNNs with attention mechanisms, enabling automatic learning of the weight distribution across various regions of the image. For instance, Ma et al. introduced FusionGAN [43], which amalgamates multi-scale convolutional networks and attention mechanisms to improve the retention of detailed information in fused images. Furthermore, AMFNet [44] merges GANs with attention mechanisms, wherein the generator employs CNNs to extract features from multi-modal images, while the attention mechanism directs the network's focus towards more salient regions within the image, thereby enhancing fusion performance. And the discriminator further augments the generator's performance by differentiating between fused and authentic images, ensuring superior visual quality and detail retention. These methods are particularly effective for fusion tasks involving rich details, as they extend the perception of global information and improve the capture of critical image features.

Transformers, in contrast, utilize self-attention mechanisms to perform global feature extraction and fusion across the entire image. This makes them especially adept at processing global information. For example, Vs et al. [45] presented the Image Fusion Transformer, illustrating the efficacy of Transformers in handling complex multi-modal images. MATR [46] highlighted the application of Transformers in multi-scale medical image fusion, capturing global features via a multi-scale adaptive mechanism. The Swin Transformer [47] manages both local and global information through a sliding window mechanism, making it well-suited for multi-modal medical image fusion. SwinFusion [48] integrates the multi-level attributes of Transformers with the sliding window mechanism, proficiently addressing complex long-range dependencies and substantially enhancing fusion performance.

While Transformer-based models are capable of capturing global contextual information through self-attention mechanisms and enhancing the overall consistency of fusion, their high computational cost constrains their widespread applicability in practical scenarios. Consequently, this study adopts the ACNN. ACNN leverages CNNs to extract local features from multi-modal images and incorporates attention mechanisms to extend the perception of global information, guiding the network's focus towards key regions within the image. Additionally, adaptive frequency-domain optimization techniques are integrated to enhance the global structure and detailed representation of the fused images, thereby significantly improving fusion performance.

3. Methodology

In this section, we first delineate the workflow of the proposed AFPNet, followed by an in-depth examination of each module's structure. Specifically, Section 3.1 provides a comprehensive overview of the framework. Section 3.2 presents the spatial domain feature extraction module, while Section 3.3 elucidates the adaptive frequency-domain weighting strategy and provides a detailed explanation of the adaptive frequency-domain weighting loss function.

3.1. Overview

AFPNet is employed for PET-MRI and SPECT-MRI image fusion. In the source images, SPECT and PET are three-channel images (i.e., RGB images), necessitating the consideration of both chrominance and luminance information, while MRI is a single-channel image containing only luminance information. Consequently, SPECT and PET images are first converted from RGB to YCbCr color space as described in Eq. (1), with their Y channels (luminance channels) fused with the corresponding MRI images. The source images input to the network are denoted as I_a with dimensions $H\times W\times C_a$ and I_b with dimensions $H\times W\times C_b$, where H and W denote height and width, respectively, and C refers the number of channels, with $C_f = \max(C_a, C_b)$. I_a with $C_a = 1$ represents the Y channel of SPECT/ PET, while I_b with $C_b = 1$ denotes the corresponding MRI image. Following the fusion process, the resulting image $I_f(C_f = 1)$ retains only luminance information. To produce a fused image containing chrominance data, $I_f(C_f = 1)$ is then combined with the color components (Cb and Cr channels) of SPECT/PET images. The final fused image $I_f(C_f = 3)$ is obtained by converting it back from YCbCr to RGB color space.

$$\begin{pmatrix} Y \\ Cb \\ Cr \end{pmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.169 & -0.331 & 0.500 \\ 0.500 & -0.419 & -0.081 \end{bmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} + \begin{pmatrix} 0 \\ 128 \\ 128 \end{pmatrix} \tag{1}$$

AFPNet is designed to optimize multi-level spatial-domain feature fusion adaptively through frequency-domain characteristics, improving the detail representation and global consistency of the fused image.As illustrated in Fig. 3, the spatial domain primarily extracts spatial features using a ACNN, progressively achieving channel domain, spatial, and multi-scale feature fusion via a three-layer fusion network. This multi-stage, multi-level feature fusion mechanism allows the model to comprehensively account for the feature information of input images across various levels and dimensions, thereby enhancing the overall performance of the fusion outcomes. Additionally, a feature reuse strategy is implemented, wherein the original input images are repeatedly reintroduced into various layers, enabling the model to leverage this information at multiple levels. This approach ensures detail retention throughout the fusion process while enhancing the expression of spatial features.

In the frequency domain, the source images' frequency characteristics are analyzed via the Fast Fourier Transform (FFT), and then adaptively employed to adjust the weights of the two source images, optimizing the network's performance in the spatial domain during the fusion process. This frequency-domain adaptive optimization enhances the fusion of functional and structural information and further improves the preservation of high-frequency details with low-frequency global

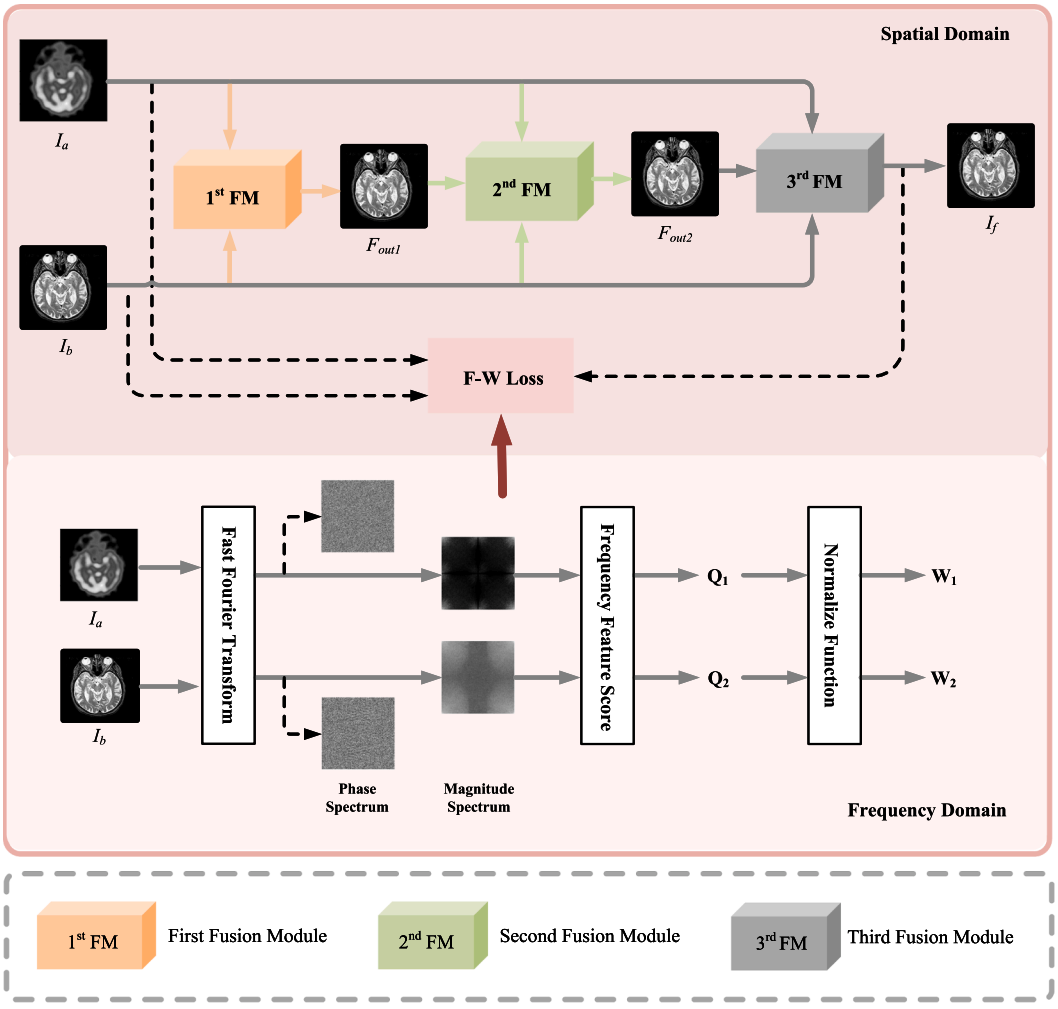


Fig. 3. Architecture of AFPNet Integrating Spatial and Frequency Domains for Enhanced Image Analysis.

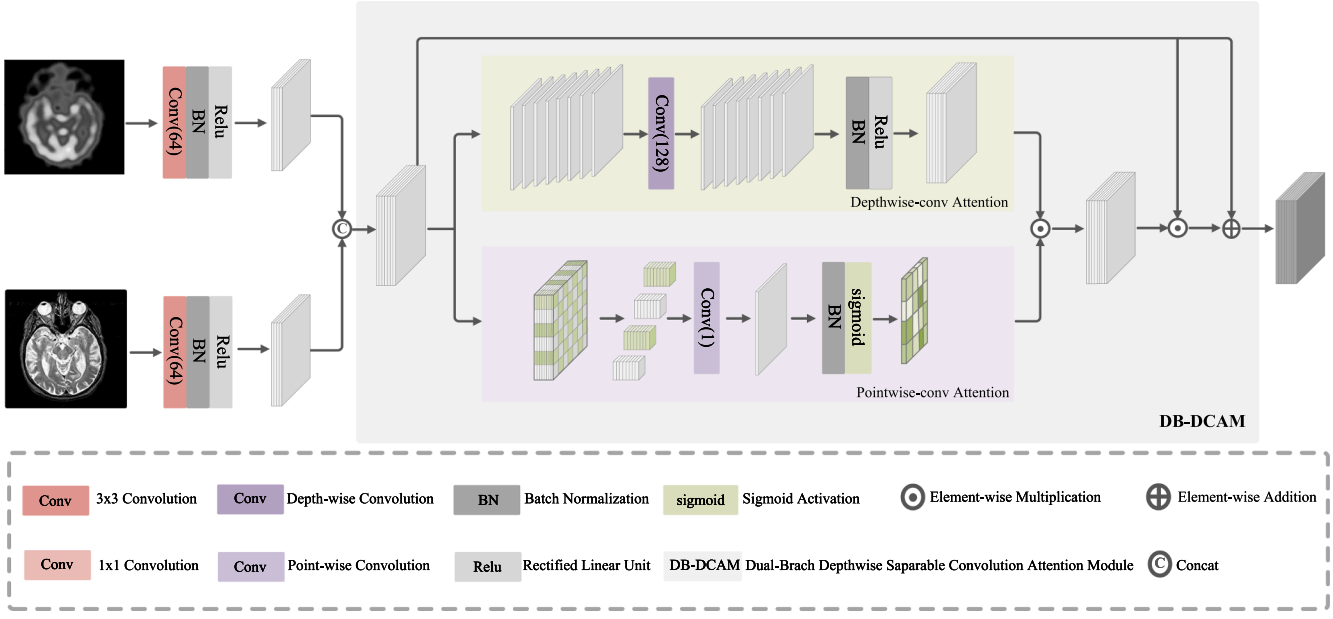


Fig. 4. The structure of the first fusion module(1st FM), with the core unit being the DB-DSCAM for fine-grained feature extraction and attention-driven fusion.

information, improving the quality of the final fused image.

3.2. HYPERLINK "SPS:id::Sec5" Spatial domain network

The Dual-Branch Depthwise Separable Convolution Attention Module (DB-DSCAM) in the initial layer meticulously processes image features with fine granularity. The depthwise convolution concentrates on the spatial intricacies of individual images, capturing detailed local features, whereas the pointwise convolution underscores the fusion of channel-specific features across multimodal images, thereby facilitating interaction and integration of features from diverse modalities. This dual approach guarantees the precise extraction of intricate information and dynamically adjusts pixel values within the feature map via the attention mechanism, accentuating critical regions and channels. Consequently, it not only preserves the original features but also optimizes the performance of the fused image, thereby enhancing both the discriminative power and the representational capacity of the features. As illustrated in Fig. 4, initially, the model processes the input images through the feature extraction module:

$$F^{i}_{out} = Re(Bn(Co^{3\times3}(Re(Bn(Co^{3\times3}(I_i)))))) \quad (2)$$

where $I_i(i = a, b)$ denote the Y channel of PET/SPECT images and their corresponding MRI counterparts, $Co^{3\times3}(\cdot)$ signifies the 3 × 3 convolution operation, $Bn(\cdot)$ signifies batch normalization, and $Re(\cdot)$ represents the ReLU activation function.

Subsequently, these preliminary features are input into the DB-DSCAM. In the upper branch, depthwise convolution processes each channel independently, ensuring comprehensive capture of spatial features within each channel and enabling the model to concentrate on significant spatial regions of each input image:

$$Dw\_Attn = Re(Bn(Conv^{3\times3}_{dw}(Cat(F^{a}_{out}, F^{b}_{out})))) \quad (3)$$

where F^{i}_{out} corresponds to the output feature from the feature extraction module of the source image $I_i(i = a,b)$, respectively. $Conv^{3\times3}_{dw}(\cdot)$ denotes depthwise separable convolution, $Bn(\cdot)$ signifies batch normalization, and $Re(\cdot)$ represents the ReLU activation function.

In contrast, the lower branch employs a pointwise convolution to reassemble feature channels, reducing the feature dimensionality to a single channel. This facilitates effective integration of information across different channels, with the resulting feature map highlighting significant feature relationships between channels:

$$Pw\_Attn = Sig(Bn(Conv^{1\times1}_{pw}(Cat(F^{a}_{out}, F^{b}_{out})))) \quad (4)$$

where F^{i}_{out} corresponds to the output feature from the feature extraction module of the source image $I_i(i = a,b)$, respectively. $Conv^{1\times1}_{pw}(\cdot)$ denotes pointwise convolution, $Bn(\cdot)$ signifies batch normalization, and $Sig(\cdot)$ represents the sigmoid activation function.

Ultimately, the output features from both branches are multiplied to generate an attention map where Pointwise Convolution Attention assigns fusion weights to each channel of Depthwise Convolution Attention, ensuring that the map contains both spatial features of each channel and and preserves essential inter-channel feature relationships. Subsequently, a residual connection further enhances fused image performance by integrating the original input features, retain the original input features while enhancing the expressive capability of the fused image, thereby the first fusion module achieving preliminary channel feature fusion:

$$F_{out1} = Cat(F^{a}_{out}, F^{b}_{out}) * (Dw\_Attn * Pw\_Attn) + Cat(F^{a}_{out}, F^{b}_{out}) \quad (5)$$

where F^{i}_{out} corresponds to the output feature from the feature extraction module of the source image $I_i(i = a, b)$, respectively. Dw_Attn, and Pw_Attn represent the depthwise convolution attention and pointwise convolution attention feature maps, respectively.

In the initial stage, the ACNN implements preliminary attention mechanisms, performing initial interactions and fusions at the channel level, with a focus on the independent processing of spatial features. Consequently, the objective of the second stage is to extract and fuse spatial feature information from the images more comprehensively. To achieve this, we introduce a Global Enhancement Attention Module (GEAM) that comprehensively fuses spatial information from multimodal images, as illustrated in Fig. 5.

Initially, the feature $F^{i}_{out}(i = a, b)$ of the source images is concatenated with the output feature F_{out1} from the first stage. This feature reuse strategy not only prevents information loss but also preserves detailed information from the source images across different scales, thereby enhancing the model's robustness to input variations and improving the comprehensive and precise performance of the fused image. The concatenated features are then subjected to additional convolutional processing:

$$F^{i}_{out} = Re(Co^{1\times1}(Re(Bn(Co^{3\times3}(Cat(F^{i}_{out}, F_{out1})))))) \quad (6)$$

where F_{out1} denotes the output feature from the first stage, F^{i}_{out} corresponds to the output feature from the feature extraction module of the source image $I_i(i = a, b)$, $Co^{3\times3}(\cdot)$ denotes the 3 × 3 convolution operation, $Co^{1\times1}(\cdot)$ denotes the 1 × 1 convolution operation, $Bn(\cdot)$ signifies batch normalization, and $Re(\cdot)$ denotes the ReLU activation function.

Subsequently, these features are fed into the GEAM, which employs global pooling operations to extract statistical characteristics from each channel, thereby more comprehensively reflecting the global spatial information of the images. Specifically, the Global Average Pooling (GAP) computes the mean value of the input image, while the Global Max Pooling (GMP) finds the maximum value for each channel. These statistical features are processed through a linear layer, integrating global spatial information from both images to to generate a global attention map, thereby enhancing multimodal image correlation and global consistency:

$$F^{i}_{max} = GMP(F^{i}_{out})$$

$$F^{i}_{avg} = GAP(F^{i}_{out}) \quad (8)$$

$$F_{out2} = Sig(linear(F^{a}_{max} + F^{b}_{max}) + linear(F^{a}_{avg} + F^{b}_{avg})) \quad (9)$$

where F^{i}_{out} corresponds to the output feature from the feature extraction module of the source image $I_i(i = a,b)$, $linear(\cdot)$ denotes the linear layer, $GMP(\cdot)$ denotes global max pooling, $GAP(\cdot)$ denotes global average pooling, F^{i}_{max} and F^{i}_{avg} represent the maximum and average values of each channel of $I_i(i = a,b)$, respectively, and $Sig(\cdot)$ denotes the sigmoid activation function.

The third layer utilizes the Multi-Scale Feature Fusion Module (MSFM), which is designed to further integrate and refine the features extracted in the preceding two stages to produce the final fused image, as illustrated in Fig. 6. Initially, the global attention map F_{out2} from the second stage is applied to the source images to generate weighted features:

$$F^{i}_{w\_out} = F_{out2} * I_i + I_i \quad (10)$$

where $I_i\ (i = a, b)$ denotes the input image, F_{out2} represents the global attention map from the second stage.

Subsequently, the weighted features $F^{i}_{w\_out}\ (i = a, b)$ are concatenated and fed into the multi-scale dilated convolution module. By employing dilated convolutions with varying dilation rates, the model captures contextual information at multiple scales and integrates this information for feature fusion. The inclusion of dilated convolution not only enlarges the receptive field but also preserves resolution, thereby more effectively capturing multi-scale features and improving the overall performance of the fused image:

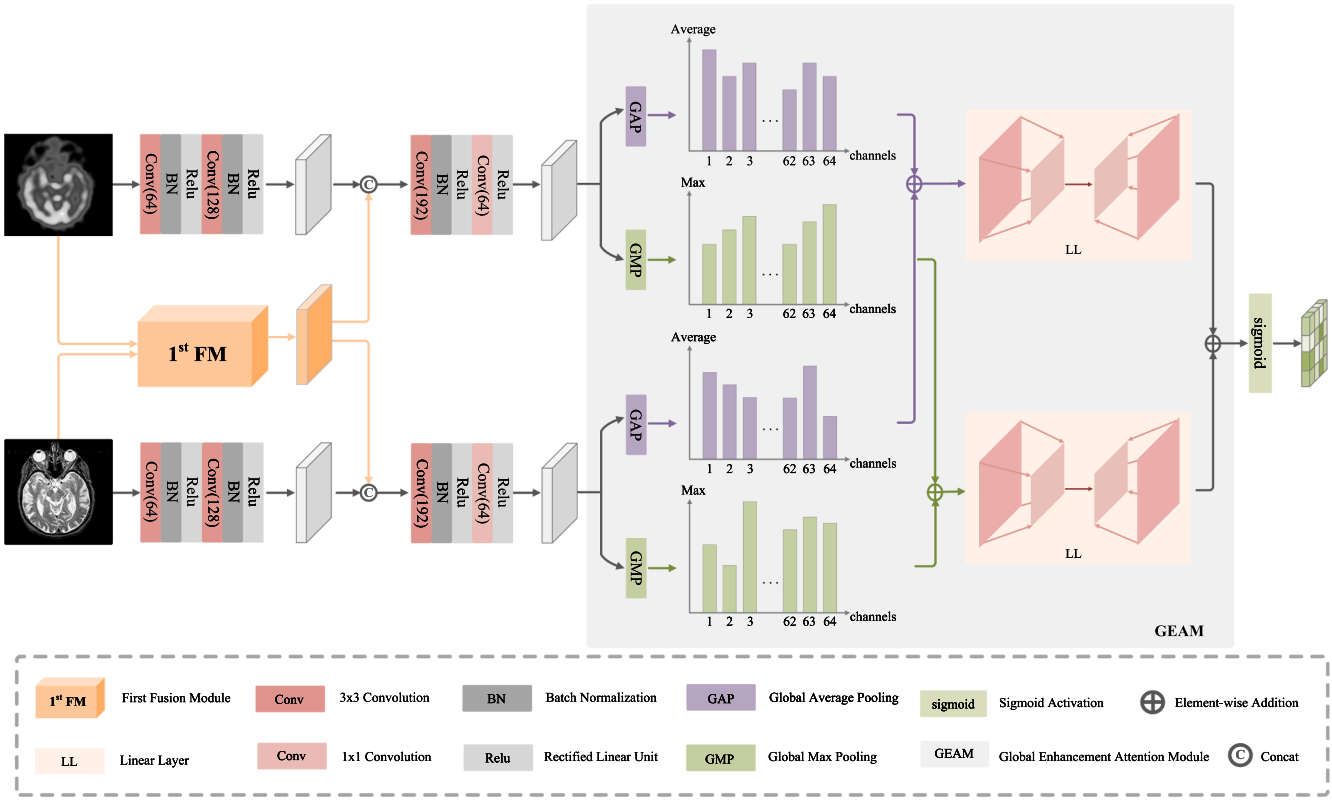


Fig. 5. The structure of the second fusion module(2nd FM), with the core unit being the GEAM for Comprehensive Spatial Feature Fusion in Multimodal Images.

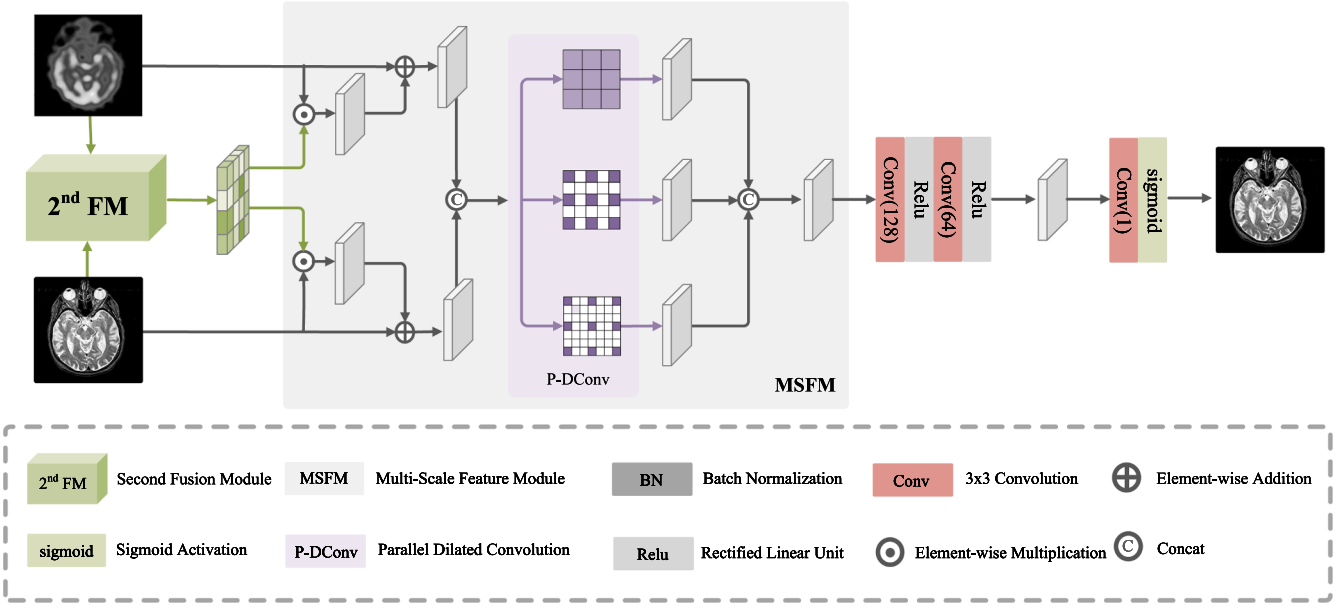


Fig. 6. The structure of the third fusion module(3rd FM), with the core unit being the MSFM for Enhanced Integration and Refinement of Multimodal Image Features.

$$F_{out}^{j} = Conv_{dilated}^{j}(Cat(F_{w\_out}^{a}, F_{w\_out}^{b})) \tag{11}$$

$$I_f = Sig(Co^{1\times1}(Re(Co^{3\times3}(Re(Co^{3\times3}(Cat(F_{out}^{1}, F_{out}^{2}, F_{out}^{3})))))) \tag{12}$$

Where $F_{w\_out}^{i}$ correspond to the weighted output features of $I_i (i = a, b)$, respectively, $Conv_{dilated}^{j}(\cdot)$ denotes the jth $(j = 1, 2, 3)$ dilated convolution, F_d^j denotes the output feature of the $j$jth dilated convolution operation, $Re(\cdot)$ denotes the ReLU activation function, and $Sig(\cdot)$ denotes the sigmoid activation function, I_f denotes the final fused image.

3.3. Adaptive frequency-domain weighted loss function

Although the progressive three-level ACNN effectively extracts both local and global image features in the spatial domain, it may still lead to information loss when handling different modalities. By contrast, frequency-domain analysis methods offer a significant advantage in

minimizing such losses, as they uncover subtle features that are less detectale in the spatial domain, thereby improving the fusion of high-frequency details and low-frequency structures. This enhances the overall quality of the fused image, ensuring better preservation of critical diagnostic information, such as tissue boundaries, edges, and other fine features.

By applying the FFT [49,50], source images are transitioned from the spatial to the frequency domain, where low-frequency components represent overall structures, and high-frequency components encapsulate finer details and edge information. For instance, MRI images predominantly capture anatomical structures with abundant detail. Conversely, PET images primarily depict the spatial distribution of physiological functions, and are characterized by coarser structural representations. Yet, an analysis of the frequency spectrum in Fig. 7 (a) reveals that despite emphasizing fine details and edges, MRI images retain substantial low-frequency components. Similarly, although PET images predominantly present macroscopic structures, frequency-domain analysis uncovers high-frequency components especially in areas with complex tissue boundaries or high functional activity. As illustrated in Fig. 7**(b)**, MRI images in the spatial domain mainly capture soft tissue details and edge information, and their frequency spectrum show significant high-frequency components, aligning with their spatial-domain fetures. In contrast, SPECT images exhibit predominantly low-frequency components in the frequency domain, indicating a limited capacity for fine detail and suitability for global structures and functional distribution presentation. global structures and functional distribution presentation. These divergences in frequency-domain characteristics form an optimal foundation for image fusion.

To overcome the limitations of spatial-domain methods in capturing information during multimodal image fusion, AFPNet proposes an adaptive frequency-domain optimization strategy. This approach extracts frequency features through frequency-domain analysis and optimizes the spatial-domain network to enhance image fusion performance. As illustrated in Fig. 3**(b)**, AFPNet applies FFT to convert source images to the frequency domain, extracting their magnitude spectrum and calculating the statistical mean to quantify the strength of each frequency components. The statistical values assessthe strength of both high and low-frequency components, capturing the comprehensive frequency characteristics of the image while preserving both essential details and broader structures. During the fusion process, AFPNet adaptively adjusts the weights ω_a and ω_b for each image based on their respective frequency domain strengths Q_a and Q_b, preserving essential both details and structural information in the fused output.

Let the input image be represented as a matrix $f(x,y)$ of size $M\times N$, where x and y are the spatial coordinates of the image, and $f(x,y)$ corresponds to the pixel value in the spatial domain. The two-dimensional discrete Fourier transform (2D DFT) is formally defined as:

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\bullet e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)} \tag{13}$$

where $F(u,v)$ is a complex value in the output frequency domain, representing the Fourier coefficient at the frequency coordinates (u,v), j is the imaginary unit, and e is the base of the natural logarithm. By applying the two-dimensional discrete Fourier transform, the spatial domain image (x,y) is transformed into its frequency domain counterpart $F(u,v)$.

$F(u,v)$ encodes both the intensity and phase information at each frequency, providing a comprehensive depiction of the image in the frequency domain. Specifically, the magnitude of $F(u,v)$ represents the amplitude at frequency (u,v), whereas its phase angle conveys the phase information. Consequently, $F(u,v)$ integrates both frequency and phase characteristics, encapsulating the energy distribution and phase correlations across the frequency spectrum.As illustrated in Fig. 3(b), within AFPNet, the extraction of the magnitude spectrum of the Fourier transform is essential to quantify the frequency feature intensity:

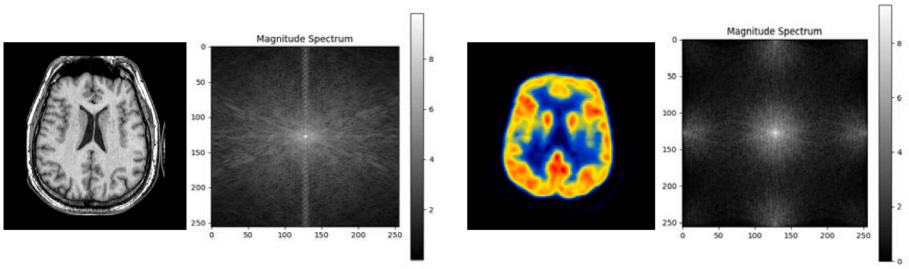
$$\text{magnitude\_spectrum}(I_i)_{u,v}=\log(|F(u,v)|+1) \tag{14}$$

Subsequently, the statistical mean of the frequency characteristics for each image is calculated according to the following formula:

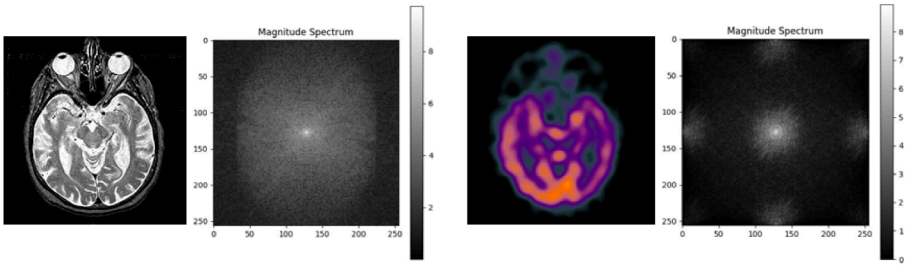
$$Q_i=\frac{1}{MN}\sum_{u=0}^{M-1}\sum_{v=0}^{N-1}\left\{\text{magnitude\_spectrum}(I_i)_{u,v}\right\} \tag{15}$$

The corresponding normalized weights are subsequently derived as follows:

$$\omega_a=Q_a^2/(Q_a^2+Q_b^2)\quad \omega_b=Q_b^2/(Q_a^2+Q_b^2) \tag{16}$$



(a) MRI-PET



(b) MRI-SPECT

Fig. 7. Frequency-Domain Analysis of MRI, PET, and SPECT Images: Comparative Assessment of Anatomical and Functional Details.

where $I_i(i = a, b)$ represents the input images, $\text{magnitude\_spectrum}(I_i)_{u,v}$ denotes the magnitude spectrum at frequency coordinates (u, v) for the input image I_i. $Q_i(i = a, b)$ represents the statistical mean of the frequency characteristics for the entire image I_i, and $\omega_i(i = a, b)$ signifies the normalized weight for the input image I_i.

Based on the above weights, AFPNet designs a frequency-domain weighted loss function that dynamically adjust the weights to accommodate frequency feature variations optimizing the spatial-domain network, enhancing the complementary information between different modalities. This dynamic adjustment mechanism allows AFPNet to more precisely fine-tune the features of the fused image, thereby enhancing the overall quality:

$$L = \omega_a L_a + \omega_b L_b \tag{17}$$

where $\omega_i(i = a, b)$ represents the normalized weight for the input image I_i. The loss terms L_a and L_b are defined by Mean Absolute Error (MAE) and Structural Similarity Index (SSIM):

$$L_i = \alpha L_{ssim} + \beta L_{mae} \tag{18}$$

where $i = a, b$, and α , β are hyperparameters used to adjust the weight of the SSIM loss and MAE loss in the total loss.

In multimodal medical image fusion, preserving important details and structural information from different modality images and reduce noise and artifacts is essential for clinical applications. AFPNet employs the MAE loss and the SSIM loss functions to achieve this goal by leveraging their respective strengths.

The MAE loss directly quantifies pixel-level discrepancies between the fused and input images,ensuring numerical fidelity and reducing noise:

$$L_{mae} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left| I_f(i, j) - I_i(i, j) \right| \tag{19}$$

where $I_i(i = a, b)$ represents the input images, I_f represents the fused image.

This is crucial for retaining metabolic information in PET/SPECT and anatomical precision in MRI. However, the MAE loss does not capture image textures, edges, and other structural features, which may result in structural inconsistencies in the fused image.To address this, the SSIM loss function is incorporated, which evaluates image similarity in terms of luminance, contrast, and structural integrity, making it highly sensitive to perceptual variations in image structure:

$$L_{ssim} = 1 - \text{SSIM}(I_f, I_i) \tag{20}$$

where $I_i(i = a, b)$ represents the input images, I_f represents the fused image.

The SSIM helps to preserve critical structural details such as organ boundaries and textures, improving the visual quality of the fused image.

In medical image fusion, this combination leads to a more accurate and clinically useful fused image.

4. Experimental settings

In this section, we outline the experimental setup for the proposed AFPNet. Specifically, Section 4.1 details the datasets and experimental configuration, while Section 4.2 introduces the comparison methods and evaluation metrics.

4.1. Datasets and experimental setup

For the two medical image fusion tasks, we utilized the widely recognized dataset "The Harvard Whole Brain Atlas",[1] from which two sets of co-registered multimodal medical image pairs were derived: MRI-PET (269 pairs) and MRI-SPECT (357 pairs). All images were standardized to dimensions of 256×256, with pixel intensity values ranging between [0, 255]. A total of 20 image pairs were randomly selected for the test set, while the remaining images were allocated to the training set.To augment the training dataset, the images were cropped into patches of size 120×120 with a stride of 20, resulting in a total of 12,201 MRI-PET and 16,513 MRI-SPECT training samples.

All experiments were conducted on a system equipped with an NVIDIA GeForce RTX 4060 Ti GPU, utilizing the PyTorch framework for implementation. The model was trained using the Adam optimizer over 35 epochs, with a batch size of 32 and an initial learning rate of 1e-3, which decayed by a factor of 0.1 after 20 epochs. The hyperparameters α and β were assigned values of 1 and 4, respectively.

4.2. Comparison methods and evaluation metrics

To evaluate the performance of AFPNet, we compared it against eight state-of-the-art deep learning-based medical image fusion (MIF) algorithms: DDcGAN [35], U2Fusion [51], PMGI [52], EMFusion [53], SDNet [54], Dilran [11], MATR [46], and FATFusion [55].

Eight widely recognized evaluation metrics were employed to quantitatively assess the fusion performance from various perspectives, including Mutual Information (MI), Visual Information Fidelity (VIF), Normalized weighted edge information (Qabf), Edge intensity(EI), Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE), and Local Mutual Information (LMI). MI metric quantifies the information content transmitted from the source images to the fused image. VIF index evaluates the fidelity of the fused image with respect to the reference image. Qabf metric is a specialized image fusion quality measure that quantifies the transfer of edge information from source to fused images. SSIM quantifies the structural similarity between images, mimicking the human visual system's assessment of quality by analyzing luminance, contrast, and structural elements. EI quantifies the average edge intensity within an image, providing a metric for evaluating the sharpness and strength of edge features. This metric serves as an indicator of texture detail and contrast in the spatial domain, offering insight into the image's structural complexity. PSNR quantifies the signal-to-noise ratio, indicating the presence of distortion in the image. MSE metric evaluates the average squared difference between the source and fused images. LMI index assesses the similarity across localized regions within the image. In the case of MSE, higher scores denote poorer fusion quality, whereas for the other six metrics, higher values signify improved fusion performance.

5. Experimental results

In this section, we present a comprehensive comparison between our proposed AFPNet and several state-of-the-art techniques for PET-MRI and SPECT-MRI image fusion. Both qualitative and quantitative analyses were performed using publicly available datasets.

5.1. Analysis of fusion results

5.1.1. Quantitative evaluation

Table 1 and Table 2 provide a comprehensive quantitative comparison of AFPNet against eight state-of-the-art MIF techniques in the context of SPECT-MRI and PET-MRI fusion tasks. Each table reports the average performance across different metrics for 20 test images, with the

[1] https://www.med.harvard.edu/AANLIB/home.html.

Table 1
Performance Comparison of AFPNet and different network structures for SPECT-MRI Fusion.

Metrics	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
DDcGAN	1.8592	0.5602	0.6889	81.2070	1.3642	30.1407	52.2913	2.2147
U2Fusion	2.0634	0.6030	0.6461	69.7855	**1.4239**	29.2520	54.1781	2.1237
PMGI	2.2852	0.7395	0.7513	82.4217	1.1252	30.5093	52.5631	2.2389
SDNet	2.1606	0.5884	0.6688	71.5831	0.9716	29.2918	57.9341	2.1645
EMFusion	2.3924	0.7421	0.7415	78.3392	1.4189	30.3545	51.0388	2.1984
dilran	2.4105	0.7310	0.6861	73.3065	1.3923	30.8757	51.2543	2.1256
MATR	1.9993	0.5405	0.3976	53.3374	0.7483	27.5923	60.6754	2.0049
FATFusion	2.5708	0.8122	0.6543	**91.0826**	1.3669	30.8346	53.2486	2.1681
AFPNet	**3.1457**	**0.9782**	**0.7882**	84.2310	1.4217	**32.7177**	**33.3584**	**2.2442**

Table 2
Performance Comparison of AFPNet and different network structures for PET-MRI Fusion.

Metrics	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
DDcGAN	1.8822	0.5985	0.5519	85.7334	1.3927	30.0542	54.5155	2.2728
U2Fusion	1.9932	0.7016	0.6224	90.6251	1.4524	30.1119	53.7150	2.2902
PAGI	2.0988	0.7362	0.6869	93.3396	0.7290	28.7251	83.7545	2.3355
SDNet	1.9569	0.6815	0.5371	79.0288	0.7597	28.6208	70.2287	2.2787
EMFusion	2.2697	0.8453	0.6819	91.9362	1.4519	30.7585	48.6186	2.2868
dilran	2.1608	0.7798	0.6488	81.8137	1.4437	30.4469	50.3116	2.2589
MATR	2.5042	0.8229	0.6505	71.9530	0.6143	28.3907	74.3686	2.2947
FATFusion	2.2954	0.8133	0.6863	**104.3312**	0.6414	28.0742	100.5286	2.3229
AFPNet	**2.7856**	**0.9635**	**0.7223**	97.1439	**1.4615**	**32.7057**	**34.6205**	**2.4000**

optimal values distinctly highlighted in bold for clarity.

It is evident from the tables that AFPNet generally performs very well across most evaluation metrics, highlighting its significant advantages in both SPECT-MRI and PET-MRI fusion. This underscores that AFPNet can effectively extract complementary information from multimodal inputs, integrating unique features and intricate details to generate more comprehensive, enriched fused images. Notably, AFPNet attains the highest scores for both MI (mutual information) and LMI (local mutual information) metrics in both sets of results. This implies that AFPNet efficiently fuses multimodal information at a global level while preserving more intricate details at the local scale. Furthermore, AFPNet's highest *Q*abf value underscores its superior ability to capture edge information from the input images. Its elevated EI score further highlights its effectiveness in preserving edge details and maintaining high-intensity textures in the fused images. The elevated SSIM score indicates that AFPNet's fused images exhibit a high degree of structural similarity to the input, preserving more of the original structure. The highest VIFscore denotes that AFPNet produces fused images with superior visual fidelity to the input, resulting in more realistic and natural outputs. The highest PSNR indicates reduced distortion in AFPNet's fused images, whereas the lowest MSE highlights minimal differences between the fused and input images, validating AFPNet's ability to reduce image distortion and retain details.

Fig. 8 and Fig. 9 illustrate the quantitative comparison between AFPNet and eight other advanced MIF methods in the SPECT-MRI and PET-MRI fusion tasks. These figures illustrate the performance metrics of individual test images, alongside their aggregated averages across the entire test dataset. In line with the table data, AFPNet generally performs outperforms all other methods across all evaluation metrics, with notable superiority in MI, LMI, *Q*abf, and EI, etc, underscoring its exceptional ability in information retention and edge extraction. The elevated SSIM and VIF scores further affirm that AFPNet's fused images demonstrate significant advantages in structural similarity and visual fidelity, whereas the PSNR and MSE results validate its ability to minimize distortion and preserve original details.

In conclusion, both the quantitative evaluations presented in the tables and the performance trends illustrated in the figures clearly demonstrate AFPNet's strengths in information retention, edge preservation, structural similarity, and overall image quality, resulting in superior fusion outcomes.

5.1.2. *Qualitative evaluation*

Fig. 10 showcases four pairs of SPECT-MRI source images and their corresponding fusion results generated by AFPNet, compared against eight other MIF techniques. AFPNet exhibits superior overall fusion performance, particularly excelling in color fidelity and detail preservation. Structural information is meticulously preserved, with sharp details and natural textures, rendering the fused images visually exceptional. In contrast, the fused outputs from DDcGAN, SDNet, and MATR appear relatively dim, displaying reduced contrast and obscured details, particularly in high-contrast regions, which diminishes visual impact and hinders complete information conveyance. U2Fusion, PAGI, and EMFusion maintain strong performance in color fidelity and detail retention, successfully preserving a substantial amount of image information. However,when processing intricate brain structures, their performance is slightly inadequate, with incomplete information fusion in certain regions, leading to a visual effect that is noticeably inferior to AFPNet. The fused images generated by Dilran and FATFusion exhibit significant color distortion, particularly in areas with complex structures, undermining both the naturalness and authenticity of the images, while diminishing detail expressiveness and overall image quality. Overall, AFPNet surpasses other methods in visual quality, detail preservation, and the naturalness and authenticity of the fused images, showcasing superior image fusion and offering more reliable visual references for medical image analysis and diagnosis.

Fig. 11 showcases comparative results between AFPNet and eight other medical image fusion methods on PET-MRI data further emphasize the exceptional performance of AFPNet.Firstly, AFPNet demonstrates exceptional overall performance throughout the fusion process. The generated images not only retain intricate details and structural fidelity but also excel in brightness, contrast, and color reproduction. AFPNet distinctly preserves edge details while ensuring the completeness of visual information, with overall quality markedly surpassing that of other methods.The fusion outputs from DDcGAN, SDNet, and MATR are generally dim, resulting in reduced image contrast and diminished visual impact. This dimness obscures critical details, especially in high-contrast regions, where these methods fail to adequately retain information, thereby affecting the overall quality of the final image. PAGI and U2Fusion perform competently in color preservation and fidelity but fall short when handling intricate structural details and edge information retention, particularly in areas with complex brain structures.

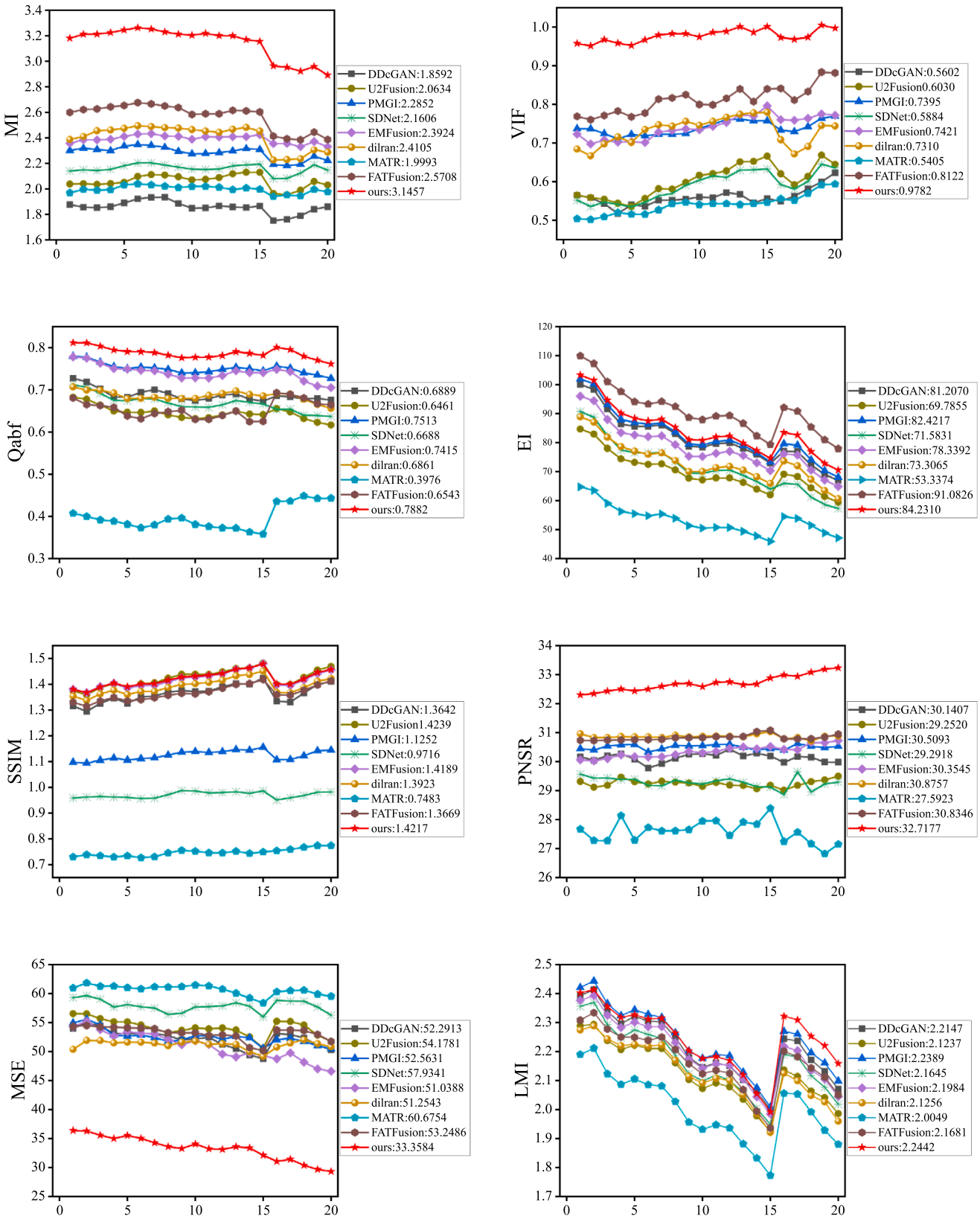


Fig. 8. Performance Comparison of Medical Image Fusion Techniques for SPECT-MRI Fusion.

These methods exhibit incomplete fusion of information in certain regions, resulting in a fusion outcome inferior to AFPNet. EMFusion demonstrates proficiency in certain areas of detail handling, successfully integrating multimodal image information. However, it falls somewhat short in color reproduction and coherence, which diminishes the overall visual consistency of the images. Dilran and FATFusion exhibit significant color distortion, particularly in regions with intricate structures, severely compromising the naturalness and authenticity of the images,

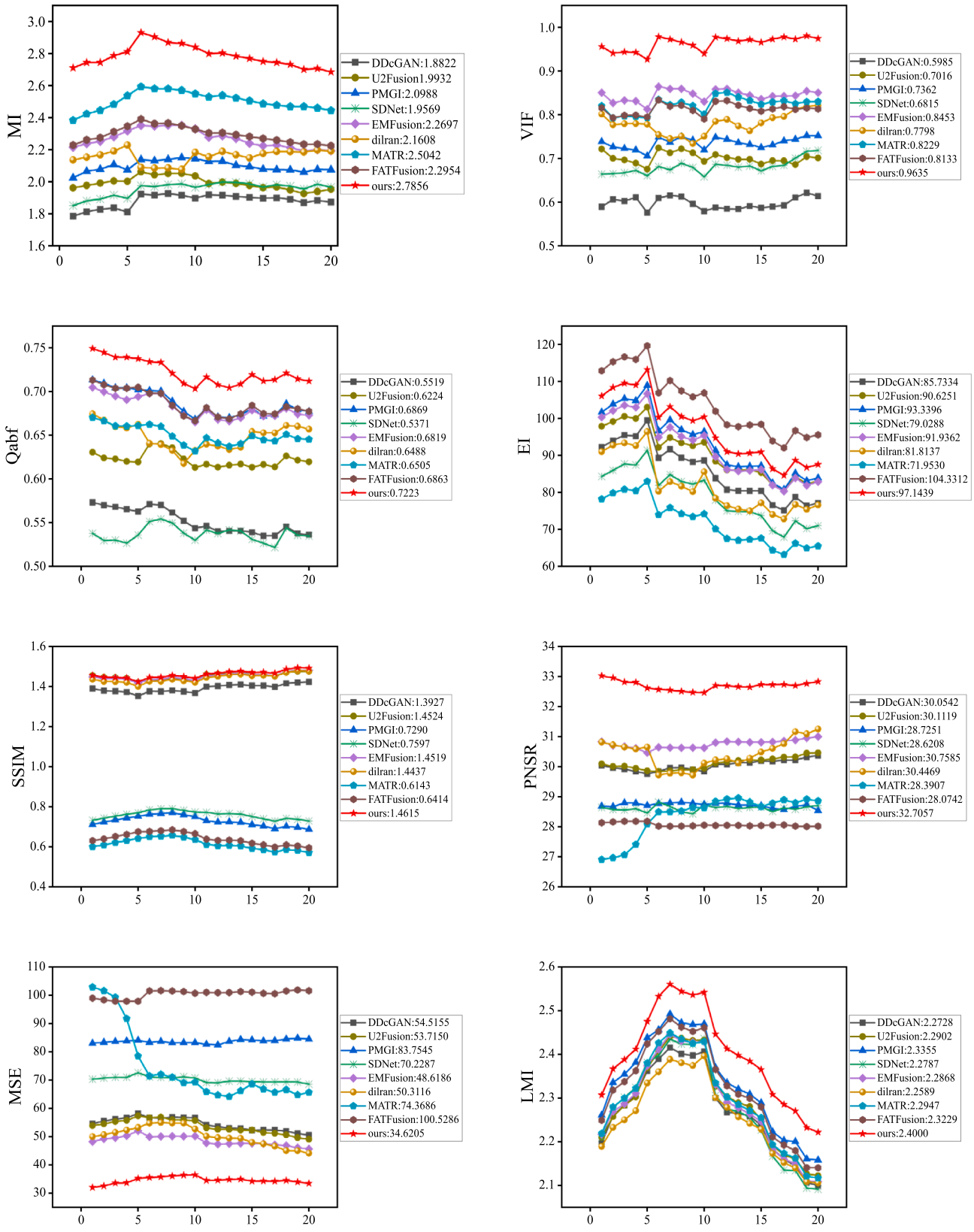


Fig. 9. Performance Comparison of Medical Image Fusion Techniques for PET-MRI Fusion.

and leading to a notable decline in overall visual quality. In summary, AFPNet outperforms other comparative methods, with its fused images not only retaining intricate detail information but also exhibiting superior color fidelity and visual coherence.

6. Ablation study

In this section, we present ablation experiments that investigate the impact of various loss functions, the frequency domain weighting

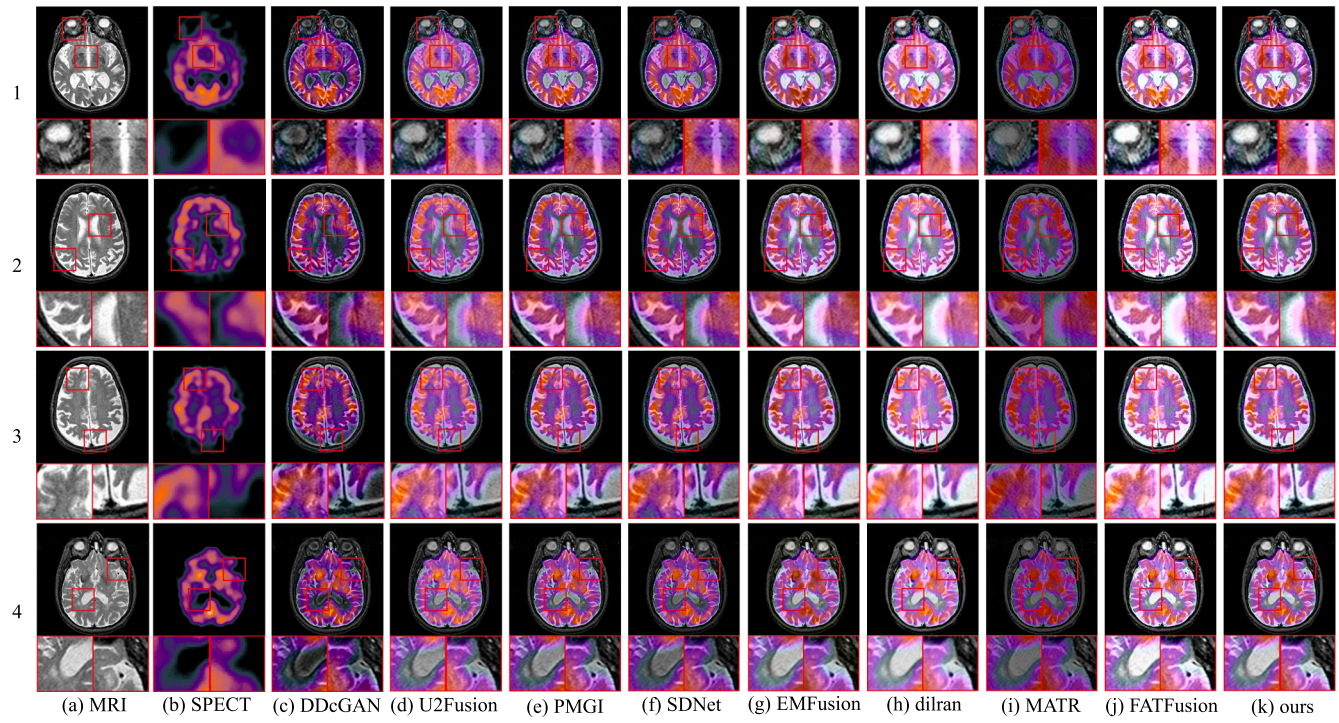


Fig. 10. Visual Comparative Fusion Results Between AFPNet and Eight Other Medical Image Fusion Techniques on SPECT-MRI Data.

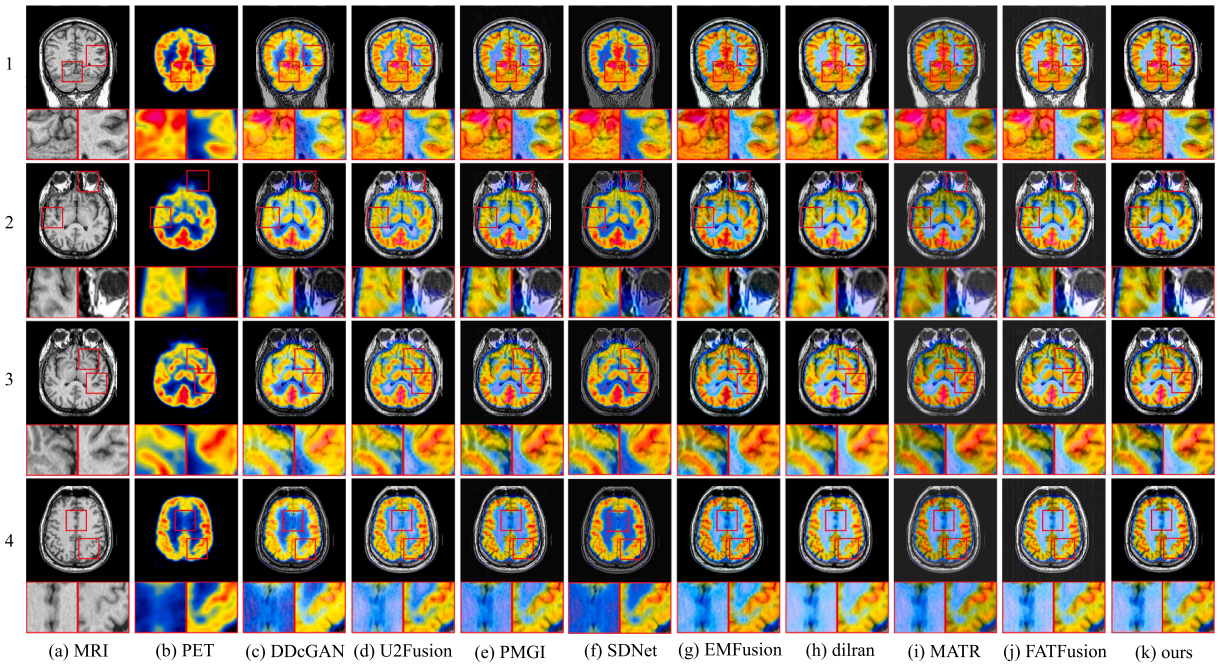


Fig. 11. Extended Comparative Fusion Results Between AFPNet and Eight Other Medical Image Fusion Techniques on PET-MRI Data.

mechanism, network structure and sequence of modules on image fusion performance. These experiments are evaluated across seven objective metrics. For clarity, the fusion results with and without the frequency domain weighting mechanism are depicted separately in Fig. 12, while the network structure ablation results are illustrated in Fig. 13.

6.1. Loss function

To investigate the necessity of L_{mae} and L_{ssim}, two additional ablation experiments were conducted: (i) w/o L_{mae}, where the MAE loss was omitted from L to demonstrate the significance of L_{mae}. (ii) w/o L_{ssim}, where the SSIM loss was removed from L to assess the importance of L_{ssim}.

Table 3 summarizes the results of the ablation experiments, emphasizing the effect of different combinations of loss functions on the fused image quality. Rmoving the MAE loss (w/o L_{mae}) caused significant declines MI, VIF, Qabf, and PSNR in the SPECT-MRI modality, which reflect inadequate information retention, while the higher MSE suggests greater errors in capturing structural variations, such as metabolic activity brightness, affecting diagnostic accuracy. Similar

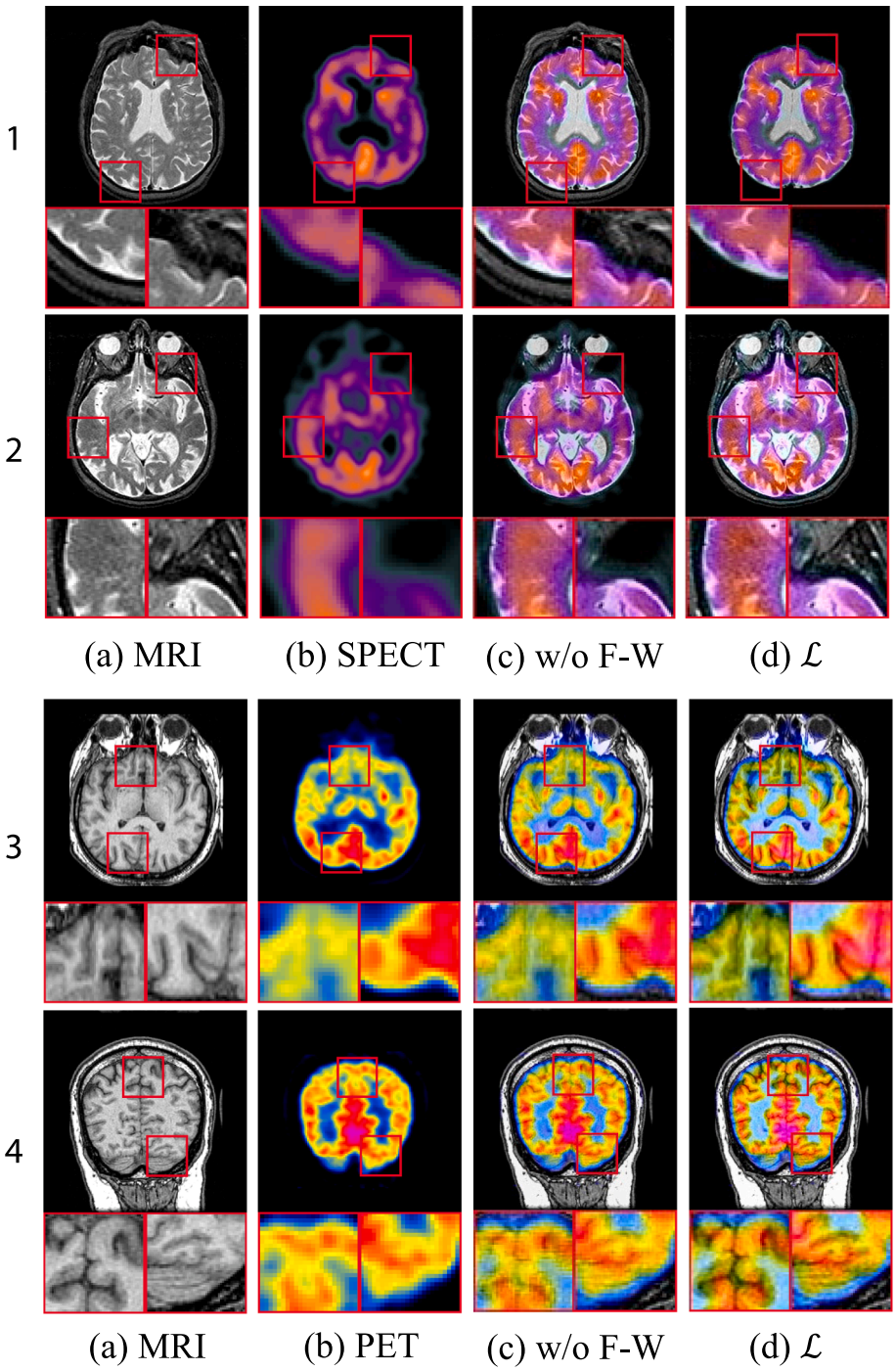


Fig. 12. Visual Comparison of Medical Image Fusion with and without the Frequency Domain Weighting Mechanism (F-W).

trends in the PET-MRI modality further underscore the importance of the MAE loss in preserving subtle details and reducing noise. Removing the SSIM loss (w/o L_{ssim}) in the SPECT-MRI modality led to declines in MI, Qabf, and LMI, impacting structural consistency and local detail preservation. Changes in PSNR, MSE, and VIF suggest reduced structural accuracy and visual quality. For the PET-MRI modality, while MI and Qabf improved, decreases in VIF, EI, PSNR, and SSIM highlight the importance of SSIM loss for structural consistency and detail retention. Overall, the SSIM loss is essential for ensuring image quality and clinical applicability.

In summary, the combination of the MAE loss and the SSIM loss ensures comprehensive optimization of the fused image, enhancing information retention, detail fidelity, and visual quality. This combination allows AFPNet to effectively retain functional information from PET/SPECT while preserving structural information from MRI, thus providing a more comprehensive and reliable imaging foundation for clinical diagnosis.

6.2. *Adaptive frequency-domain weighted loss function*

To investigate the necessity of the frequency domain weighting mechanism (F-W), additional ablation experiments were performed: w/o F-W, where the F-W was excluded from L to illustrate its significance.

The ablation experiments on the F-W further emphasize the crucial role of frequency-domain weighting mechanism in enhancing the quality of fused images. As demonstrated in Table 4, the removing of the

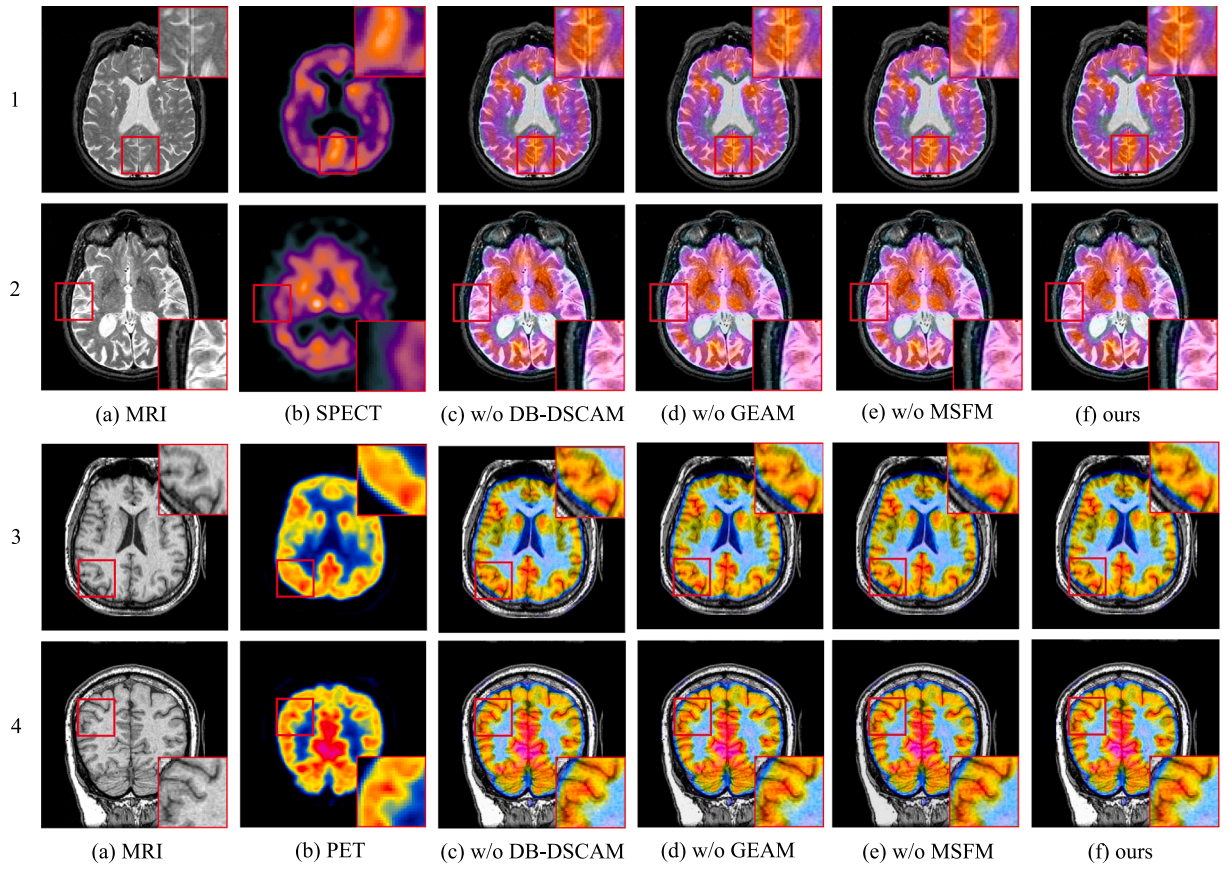


Fig. 13. Visual Impact of the DB-DSCAM, GEAM, and MSFM on Image Fusion Quality in AFPNet.

Table 3
Impact of Loss Function Combinations on Fusion Quality for SPECT-MRI and PET-MRI Modalities.

Modality	Experiments	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
SPECT-MRI	w/o L_{mae}	2.7960	0.8885	0.7674	82.1212	**1.4400**	32.0043	39.0744	**2.2767**
	w/o L_{ssim}	3.1300	0.9751	0.7871	82.9073	1.4178	32.4443	34.9046	2.2425
	L	**3.1457**	**0.9782**	**0.7882**	**84.2310**	1.4217	**32.7177**	**33.3584**	2.2442
PET-MRI	w/o L_{mae}	2.4859	0.8758	0.6955	97.5101	**1.4806**	32.1113	39.3864	2.3869
	w/o L_{ssim}	**2.8314**	0.9534	**0.7250**	95.6846	1.4558	32.6677	34.9247	**2.4037**
	L	2.7856	**0.9635**	0.7223	**97.1439**	1.4615	**32.7057**	**34.6205**	2.4000

Table 4
Effects of Frequency Domain Weighting Mechanism on Image Fusion Quality.

Modality	Experiments	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
SPECT-MRI	w/o F-W	2.1100	0.6225	0.4587	54.3400	1.4513	30.3711	49.1468	1.9021
	L	**3.1457**	**0.9782**	**0.7882**	**84.2310**	**1.4217**	**32.7177**	**33.3584**	**2.2442**
PET-MRI	w/o F-W	2.2564	0.8046	0.6246	91.4689	1.4946	31.0716	48.0494	2.3301
	L	**2.7856**	**0.9635**	**0.7223**	**97.1439**	**1.4615**	**32.7057**	**34.6205**	**2.4000**

frequency-domain weighting mechanism (w/o F-W) caused a significant decline in key performance metrics of the fused images, especially in LMI, Qabf, EI, and VIF, highlighting the substantial impact of F-W on detail retention, contrast enhancement, and edge information representation. Moreover, the deterioration of MI, PSNR, and MSE metrics indicates negative effects on the accuracy of the image's overall structure and global information retention.Fig. 12 further illustrates the essential role of the F-W in medical image fusion. The images with the F-W mechanism exhibit superior overall quality, especially in terms of high-frequency detail. Enlarged images exhibit higher contrast and sharper edges between tissues, further emphasizing the role of the F-W mechanism in preserving fine tissue structures and boundary information, both of which are crucial for medical diagnosis. In contrast, the images without the F-W mechanism exhibit noticeable blurriness, poorly defined tissue edges, loss of key structural details, unnatural color transitions, and substantially reduced contrast between tissues—deficienciesparticularly evident in complex structures such as the cerebral cortex and ventricles.

The collective results of the ablation experiments demonstrate that integrating the MAE and SSIM loss functions, together with the frequency domain weighting mechanism, is critical for enhancing the quality of medical image fusion. The combination of the complete loss function configuration and the frequency domain weighting mechanism enables AFPNet to excel across diverse modalities, significantly

improving the overall quality of the fused images.

6.3. Network structure

The core components of the proposed Spatial Domain Network consist of three critical modules: DB-DSCAM, GEAM, and MSFM. To evaluate the significance of each component, a series of ablation experiments were conducted on the validation set: (i) w/o DSCAM: Only the second and third stages were retained to assess the role of DSCAM. (ii) w/o GEAM: Only the first and third stages were kept to evaluate the role of GEAM. (iii) w/o MSFM: The first and second stages, together with the 3×3 convolution, were maintained to assess the importance of MSFM.

Table 5 presents the ablation experiment results, providing a detailed analysis of each module's contribution to the overall quality of image fusion in the AFPNet model. The experimental results demonstrate that the contributions of each module to the fused images vary significantly. Specifically, the DB-DSCAM plays a pivotal role in extracting fine details and integrating channel features. The removal of this module results in a marked decline in VIF and Qabf metrics, underscoring its essential role in ensuring detail fidelity and channel feature integration. The GEAM excels at enhancing the global information processing capabilities of the images. Following the removal of GEAM, PSNR and MSE metrics deteriorated significantly, indicating that GEAM plays a crucial role in maintaining the consistency of the overall structure and contrast, with particular emphasis on global information fusion. The MSFM, by extracting and fusing multi-scale features, further enhances detail representation and preserves high-frequency information within the images. Upon removal of this module, the SSIM and PSNR metrics for SPECT-MRI and PET-MRI images declined significantly, indicating that the MSFM is critical for capturing contextual information at multiple scales, particularly in handling high-frequency details and texture information. In conclusion, the AFPNet model's modules, through their synergistic effects, perform exceptionally well across various evaluation metrics, thereby validating the overall design's effectiveness.

Fig. 13 further substantiates the critical contribution of each module to the AFPNet model's image fusion quality. Images with the DB-DSCAM module exhibit sharper edges and finer details, particularly in enlarged regions. Removing this module results in blurred details and less sharp edge transitions compared to the complete model. This indicates that the DB-DSCAM module plays a crucial role in refining image details and sharpening edges. Removing the GEAM module reduces overall contrast, affecting brightness in brain structures, and the enlarged regions show imbalances in brightness and contrast, reducing image consistency. This corresponds with the significant changes in PSNR and MSE, further demonstrating the essential role of the GEAM module in maintaining global information consistency. After the removal of the MSFM module, the image appears relatively flat and lacking in depth, which affects its overall detail representation. Images retaining the MSFM exhibit richer detail, with smaller structures clearly visible. This phenomenon is particularly evident in the enlarged regions and corresponds with decreases in SSIM and PSNR, validating the MSFM module's importance in multi-scale feature fusion and preserving high-frequency information. Comparing the fused images from different experiments, it is evident that those with all modules retained exhibit optimal sharpness, contrast, and detail. Removing any module degrades image quality, especially in detail preservation and global contrast.

To further explore the contributions of the DB-DSCAM, GEAM, and MSFM modules in information extraction and fusion, AFPNet performed additional ablation experiments by substituting the removed modules with standard 3×3 convolutional layers. As presented in Table 6, the fused images exhibited significantly degraded quality metrics. This highlights the irreplaceable role of the DB-DSCAM, GEAM, and MSFM modules, particularly in their unique ability to extract and integrate global and local information effectively.

Overall, the visual analysis and quantitative metrics consistently corroborate the essential role of the DB-DSCAM, GEAM, and MSFM modules in image fusion. The exclusion of each module significantly diminishes image quality, confirming the indispensable role of these components in the AFPNet model.

6.4. Sequence of modules

To evaluate the effectiveness of the multi-stage, multi-level progressive fusion mechanism, we analyzed the influence of module sequencing on experimental results. In AFPNet, the DB-DSCAM, GEAM, and MSFM modules are arranged according to their specific roles in the fusion process. The DB-DSCAM module in the first fusion module (1st FM) performs channel information fusion and extracts spatial features of each modality independently, the GEAM module in the second fusion module (2nd FM), refines the global integration of spatial information, the MSFM module in the third fusion module (3rd FM) consolidates the outputs of the preceding layers using multi-scale feature fusion. The outputs of the first two layers lay a crucial foundation for the final multi-scale fusion. Therefore, we specifically investigated the effect of rearranging the first and second fusion modules (1st FM and 2nd FM) on the experimental results.

As presented in Table 7, the results reveal that changing the sequence of the DB-DSCAM and GEAM modules significantly degrades model performance across all metrics, highlighting the importance of the module sequence for effective fusion. In the current architecture, the DB-DSCAM module serves as the initial stage, ensuring that the features of each input image are distinctly represented. Subsequently, the GEAM module enhances global consistency and the collaborative representation of modality features. However, if GEAM is placed first, its global fusion operation interferes with the subsequent DB-DSCAM's ability to extract unique modality-specific information, thereby degrading the quality of the fusion results.Consequently, the designated module sequence optimally preserves modality-specific information and global coherence, enabling superior image fusion performance.

7. Conclusion

In this study, AFPNet achieves efficient fusion of global structure and local details through dual processing in the spatial and frequency domains. In the spatial domain, AFPNet primarily utilizes ACNN for spatial feature extraction, progressively achieving channel, spatial, and multi-scale feature fusion through a three-layer fusion network.

Table 5
Effects of Removing DB-DSCAM, GEAM, and MSFM Modules on AFPNet Fusion Quality for SPECT-MRI and PET-MRI Modalities.

Modality	Experiments	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
SPECT-MRI	w/o DB-DSCAM	3.0538	0.9445	0.7830	81.5391	**1.4260**	32.5050	34.6331	**2.2535**
	w/o GEAM	3.0535	0.9404	0.7820	83.1567	1.4179	32.7114	34.3817	2.2330
	w/o MSFM	**3.1482**	0.9209	0.7634	81.4605	1.4142	31.9148	34.9687	2.2412
	AFPNet	3.1457	**0.9782**	**0.7882**	**84.2310**	1.4217	**32.7177**	**33.3584**	2.2442
PET-MRI	w/o DB-DSCAM	2.7129	0.9430	0.7181	96.2604	**1.4640**	32.5685	35.7287	2.3902
	w/o GEAM	2.6118	0.9099	0.7159	95.4834	1.4638	31.5116	44.4622	2.3854
	w/o MSFM	2.5755	0.8964	0.7075	93.1852	1.4619	31.1558	41.1760	2.3511
	AFPNet	**2.7856**	**0.9635**	**0.7223**	**97.1439**	1.4615	**32.7057**	**34.6205**	**2.4000**

Table 6
Effects of Replacing DB-DSCAM, GEAM, and MSFM Modules with Standard Convolutional Layers on Fusion Quality for SPECT-MRI and PET-MRI Modalities.

Modality	Experiments	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
SPECT-MRI	DB-DSCAM → Conv	3.1400	0.9724	**0.7904**	81.3041	1.4246	32.4895	34.6950	2.2513
	GEAM → Conv	3.0100	0.9572	0.7866	82.3122	**1.4259**	32.1411	39.0493	2.2619
	MSFM → Conv	3.1262	0.9581	0.7861	83.2545	1.4217	31.8861	39.3507	**2.2755**
	AFPNet	**3.1457**	**0.9782**	0.7882	**84.2310**	1.4217	**32.7177**	**33.3584**	2.2442
PET-MRI	DB-DSCAM → Conv	2.7725	0.9528	0.7193	95.2116	1.4619	32.5453	35.9190	2.3898
	GEAM → Conv	2.6107	0.9036	0.7152	**97.7846**	**1.4625**	32.0866	39.9341	2.3710
	MSFM → Conv	2.7787	0.9600	0.7221	95.7813	1.4614	32.6225	35.1475	**2.4044**
	AFPNet	**2.7856**	**0.9635**	**0.7223**	97.1439	1.4615	**32.7057**	**34.6205**	2.4000

Table 7
Effects of Frequency Domain Weighting Mechanism on Image Fusion Quality for SPECT-MRI and PET-MRI Modalities.

Modality	Experiments	MI	VIF	Qabf	EI	SSIM	PNSR	MSE	LMI
SPECT-MRI	2nd → 1st	3.0653	0.9568	0.7860	83.4177	0.8937	31.9611	41.0815	**2.2667**
	AFPNet	**3.1457**	**0.9782**	**0.7882**	**84.2310**	**1.4217**	**32.7177**	**33.3584**	2.2442
PET-MRI	2nd → 1st	2.6636	0.9296	0.7160	93.6517	1.4610	32.1742	38.5718	2.3698
	AFPNet	**2.7856**	**0.9635**	**0.7223**	**97.1439**	**1.4615**	**32.7057**	**34.6205**	**2.4000**

Additionally, a feature reuse strategy ensures detail preservation and global correlation throughout the fusion process.

AFPNet employs an adaptive frequency-domain weighting mechanism to dynamically adjust the contribution of each input image during the fusion process. This approach not only optimizes the spatial domain but also enhances local detail preservation and ensures global structural consistency, thereby delivering high-quality image details and diagnostic insights at both macro and micro levels.SSIM emphasizes structural fidelity and perceptual quality, while MAE minimizes pixel-level error. AFPNet combines these two losses and employs frequency-domain weighting optimization to construct a dynamic loss function, preserving overall structural information while preventing detail loss, thereby significantly enhancing fused image quality. This multi-level spatial and frequency domain information fusion mechanism, combined with adaptive frequency-domain weight adjustment and dynamic loss function optimization, jointly contributes to AFPNet's notable improvements in image quality. Experimental results show that AFPNet exhibits clear advantages over other fusion methods in preserving overall structural information and intricate image details.Future research will focus on refining the fusion strategy through the incorporation of advanced frequency-domain analysis techniques and extending the evaluation framework to cover a more diverse range of datasets, thereby enhancing the generalizability and robustness of the proposed model.

CRediT authorship contribution statement

Dangguo Shao: Writing – original draft, Validation, Supervision, Formal analysis, Conceptualization. **Hongjuan Yang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Funding acquisition, Conceptualization. **Lei Ma:** Resources, Investigation, Conceptualization. **Sanli Yi:** Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: [MaLei reports statistical analysis and writing assistance were provided by Kunming University of Science and Technology. MaLei reports a relationship with Kunming University of Science and Technology that includes: employment and funding grants. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper].

Data availability

The authors do not have permission to share data.

References

[1] A.P. James, B.V. Dasarathy, Medical image fusion: a survey of the state of the art, Inf. Fusion 19 (2014) 4–19, https://doi.org/10.1016/j.inffus.2013.12.002.
[2] P. Kavita, D.R. Alli, A.B. Rao, Study of image fusion optimization techniques for medical applications, Int. J. Cognit. Comput. Eng. 3 (2022) 136–143, https://doi.org/10.1016/j.ijcce.2022.01.012.
[3] J. Nunez, X. Otazu, O. Fors, et al., Multiresolution-based image fusion with additive wavelet decomposition, IEEE Trans. Geosci. Remote Sens. 37 (3) (1999) 1204–1211, https://doi.org/10.1109/36.763274.
[4] A. Maćkiewicz, W. Ratajczak, Principal components analysis (PCA), Comput. Geosci. 19 (3) (1993) 303–342, https://doi.org/10.1016/0098-3004(93)90090-R.
[5] P. Chai, X. Luo, Z. Zhang, Image fusion using quaternion wavelet transform and multiple features, IEEE Access 5 (2017) 6724–6734, https://doi.org/10.1109/ACCESS.2017.2685178.
[6] A. Baghaie, S. Schnell, A. Bakhshinejad, et al., Curvelet transform-based volume fusion for correcting signal loss artifacts in time-of-flight magnetic resonance angiography data, Comput. Biol. Med. 99 (2018) 142–153, https://doi.org/10.1016/j.compbiomed.2018.06.008.
[7] Y. Liu, X. Chen, R.K. Ward, et al., Medical image fusion via convolutional sparsity based morphological component analysis, IEEE Signal Process Lett. 26 (3) (2019) 485–489, https://doi.org/10.1109/LSP.2019.2895749.
[8] J. -H. Jacobsen, J. Van Gemert, Z. Lou and A. W. M. Smeulders, "Structured Receptive Fields in CNNs," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2610-2619. Doi: 10.1109/CVPR.2016.286.
[9] L. Zhao, R. Yang, B. Yan, et al., DGFusion: an effective dynamic generalizable network for infrared and visible image fusion, Infrared Phys. Technol. (2024) 105495, https://doi.org/10.1016/j.infrared.2024.105495.
[10] B. Li, J.N. Hwang, Z. Liu, et al., PET and MRI image fusion based on a dense convolutional network with dual attention, Comput. Biol. Med. 151 (2022) 106339, https://doi.org/10.1016/j.compbiomed.2022.106339.
[11] Zhou M, Xu X, Zhang Y. An attention-based multi-scale feature learning network for multimodal medical image fusion[J]. arXiv preprint arXiv:2212.04661, 2022. Doi: 10.48550/arXiv.2212.04661.
[12] H. Xu, J. Yuan, J. Ma, MURF: mutually reinforcing multi-modal image registration and fusion, IEEE Trans. Pattern Anal. Mach. Intell. 45 (10) (Oct. 2023) 12148–12166, https://doi.org/10.1109/TPAMI.2023.3283682.
[13] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, Z. Luo, ReCoNet: Recurrent Correction Network for Fast and Efficient Multi-modality Image Fusion, in: S. Avidan, G. Brostow, M. Cissé, G.M. Farinella, T. Hassner (Eds.), Computer Vision – ECCV 2022. ECCV 2022. Lecture Notes in Computer Science, Springer, Cham, 2022, https://doi.org/10.1007/978-3-031-19797-0_31.
[14] Y. Liu, C. Yu, J. Cheng, et al., MM-Net: A mixformer-based multi-scale network for anatomical and functional image fusion, IEEE Trans. Image Process. 33 (2024) 2197–2212, https://doi.org/10.1109/TIP.2024.3374072.
[15] S. Li, S. Zhao, Y. Zhang, et al., Source-free unsupervised adaptive segmentation for knee joint MRI, Biomed. Signal Process. Control 92 (2024) 106028, https://doi.org/10.1016/j.bspc.2024.106028.
[16] J. Hong, Y.D. Zhang, W. Chen, Source-free unsupervised domain adaptation for cross-modality abdominal multi-organ segmentation, Knowl.-Based Syst. 250 (2022) 109155, https://doi.org/10.1016/j.knosys.2022.109155.

[17] J. Hong, S.C.H. Yu, W. Chen, Unsupervised domain adaptation for cross-modality liver segmentation via joint adversarial learning and self-learning, Appl. Soft Comput. 121 (2022) 108729, https://doi.org/10.1016/j.asoc.2022.108729.
[18] Z. Wan, Q. Yu, W. Dai, et al., Data generation for enhancing EEG-based emotion recognition: extracting time-invariant and subject-invariant components with contrastive learning, IEEE Trans. Consum. Electron. (2024), https://doi.org/10.1109/TCE.2024.3414154.
[19] W. Huang, X. Li, Evaluation of focus measures in multi-focus image fusion, Pattern Recogn. Lett. 28 (4) (2007) 493–500, https://doi.org/10.1016/j.patrec.2006.09.005.
[20] H. Yin, Y. Li, Y. Chai, et al., A novel sparse-representation-based multi-focus image fusion approach, Neurocomputing 216 (2016) 216–229, https://doi.org/10.1016/j.neucom.2016.07.039.
[21] Yuri Zhang, "A new automatic approach for effectively fusing Landsat 7 as well as IKONOS images," IEEE International Geoscience and Remote Sensing Symposium, Toronto, ON, Canada, 2002, pp. 2429-2431 vol.4. Doi: 10.1109/IGARSS.2002.1026567.
[22] T. Zhou, Q.R. Cheng, H.L. Lu, et al., Deep learning methods for medical image fusion: A review, Comput. Biol. Med. 160 (2023) 106959, https://doi.org/10.1016/j.compbiomed.2023.106959.
[23] M.D. Levine, A.M. Nazif, Dynamic measurement of computer generated image segmentations, IEEE Trans. Pattern Anal. Mach. Intell. 7 (2) (1985) 155–164, https://doi.org/10.1109/TPAMI.1985.4767640.
[24] C.O. Ancuti, C. Ancuti, C. De Vleeschouwer, et al., Single-scale fusion: an effective approach to merging images, IEEE Trans. Image Process. 26 (1) (2016) 65–78, https://doi.org/10.1109/TIP.2016.2621674.
[25] S. Li, J.T. Kwok, Y. Wang, Using the discrete wavelet frame transform to merge Landsat TM and SPOT panchromatic images, Inf. Fusion 3 (1) (2002) 17–23, https://doi.org/10.1016/S1566-2535(01)00037-9.
[26] Z. Wang, X. Li, H. Duan, et al., Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform, Expert Syst. Appl. 171 (2021) 114574, https://doi.org/10.1016/j.eswa.2021.114574.
[27] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (2015) 72–84, https://doi.org/10.1016/j.inffus.2014.10.004.
[28] Z. Wang, Z. Cui, Y. Zhu, Multi-modal medical image fusion by Laplacian pyramid and adaptive sparse representation, Comput. Biol. Med. 123 (2020) 103823, https://doi.org/10.1016/j.compbiomed.2020.103823.
[29] Y. Liu, X. Chen, J. Cheng and H. Peng, "A medical image fusion method based on convolutional neural networks," 2017 20th International Conference on Information Fusion (Fusion), Xi'an, China, 2017, pp. 1-7. Doi: 10.23919/ICIF.2017.8009769.
[30] A. Raza, H. Huo, T. Fang, PFAF-Net: Pyramid feature network for multimodal fusion, IEEE Sens. Lett. 4 (12) (2020) 1–4, https://doi.org/10.1109/LSENS.2020.3041585.
[31] W. Li, X. Peng, J. Fu, et al., A multiscale double-branch residual attention network for anatomical–functional medical image fusion, Comput. Biol. Med. 141 (2022) 105005, https://doi.org/10.1016/j.compbiomed.2021.105005.
[32] L. Guo, D. Tang, Infrared and visible image fusion using a generative adversarial network with a dual-branch generator and matched dense blocks, SIViP 17 (5) (2023) 1811–1819, https://doi.org/10.1007/s11760-022-02392-z.
[33] Y. Liu, X. Chen, H. Peng, et al., Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207, https://doi.org/10.1016/j.inffus.2016.12.001.
[34] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, D. Chisholm, SEDRFuse: a symmetric encoder–decoder with residual block network for infrared and visible image fusion, IEEE Trans. Instr. Measure. 70 (2021) 1–15, https://doi.org/10.1109/TIM.2020.3022438, 5002215.
[35] J. Ma, H. Xu, J. Jiang, et al., DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995, https://doi.org/10.1109/TIP.2020.2977573.
[36] J. Huang, Z. Le, Y. Ma, et al., MGMDcGAN: Medical image fusion using multi-generator multi-discriminator conditional generative adversarial network, IEEE Access 8 (2020) 55145–55157, https://doi.org/10.1109/ACCESS.2020.2982016.
[37] X. Li, M. Li, P. Yan, et al., Deep learning attention mechanism in medical image analysis: Basics and beyonds, Int. J. Network Dyn. Intell. (2023) 93–116, https://doi.org/10.53941/ijndi0201006.
[38] S. Wang, Y. Chen, Z. Yi, A multi-scale attention fusion network for retinal vessel segmentation, Appl. Sci. 14 (7) (2024) 2955, https://doi.org/10.3390/app14072955.
[39] H. Zhu, J. Wang, S.H. Wang, et al., An evolutionary attention-based network for medical image classification, Int. J. Neural Syst. 33 (03) (2023) 2350010, https://doi.org/10.1142/S0129065723500107.
[40] J. Wang, S.Y. Lu, S.H. Wang, et al., RanMerFormer: Randomized vision transformer with token merging for brain tumor classification, Neurocomputing 573 (2024) 127216, https://doi.org/10.1016/j.neucom.2023.127216.
[41] F. Shamshad, S. Khan, S.W. Zamir, et al., Transformers in medical imaging: A survey, Med. Image Anal. 88 (2023) 102802, https://doi.org/10.1016/j.media.2023.102802.
[42] X. Xie, X. Zhang, X. Tang, et al., MACTFusion: lightweight cross transformer for adaptive multimodal medical image fusion, IEEE J. Biomed. Health Inform. (2024), https://doi.org/10.1109/JBHI.2024.3391620.
[43] J. Ma, W. Yu, P. Liang, et al., FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26, https://doi.org/10.1016/j.inffus.2019.09.002.
[44] J. Wang, L. Yu, S. Tian, et al., AMFNet: An atten tion-guided generative adversarial network for multi-modal image fusion, Biomed. Signal Process. Control 78 (2022) 103990, https://doi.org/10.1016/j.bspc.2022.103990.
[45] V. Vs, J. M. Jose Valanarasu, P. Oza and V. M. Patel, "Image Fusion Transformer," 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 2022, pp. 3566-3570. Doi: 10.1109/ICIP46576.2022.9897280.
[46] W. Tang, F. He, Y. Liu, et al., MATR: Multimodal medical image fusion via multiscale adaptive transformer, IEEE Trans. Image Process. 31 (2022) 5134–5149, https://doi.org/10.1109/TIP.2022.3172625.
[47] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 2021, pp. 9992-10002.Doi: 10.1109/ICCV48922.2021.00986.
[48] J. Ma, L. Tang, F. Fan, et al., SwinFusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sin. 9 (7) (2022) 1200–1217, https://doi.org/10.1109/JAS.2022.105686.
[49] R.N. Bracewell, The fourier transform, Sci. Am. 260 (6) (1989) 86–95, https://doi.org/10.1038/scientificamerican0689-86.
[50] P. Heckbert, Fourier transforms and the fast Fourier transform (FFT) algorithm, Computer Graphics 1995 (2) (1995) 15–463.
[51] H. Xu, J. Ma, J. Jiang, et al., U2Fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2020) 502–518, https://doi.org/10.1109/TPAMI.2020.2993955.
[52] Zhang H, Xu H, Xiao Y, et al. Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12797-12804. Doi: 10.1609/aaai.v34i07.12797.
[53] H. Xu, J. Ma, EMFusion: An unsupervised enhanced medical image fusion network, Inf. Fusion 76 (2021) 177–186, https://doi.org/10.1016/j.inffus.2021.06.007.
[54] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785, https://doi.org/10.1007/s11263-021-01501-8.
[55] W. Tang, F. He, FATFusion: A functional–anatomical transformer for medical image fusion, Inf. Process. Manag. 61 (4) (2024) 103687, https://doi.org/10.1016/j.ipm.2024.103687.