

Article

DDFNet: A Dual-Stream Decoupled Feature Alignment Network for Multimodal Medical Image Fusion

Pengquan Han , Manyuan Cheng, Cuiyin Liu * and Bo Liang 

Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650000, China; hanpengquan@stu.kust.edu.cn (P.H.); chengmanyuan1@stu.kust.edu.cn (M.C.); liangbo@kust.edu.cn (B.L.)

* Correspondence: liucuiyin@kust.edu.cn

Abstract

Multimodal Medical Image Fusion (MMIF) aims to exploit the correlation and complementarity of different imaging modalities to integrate cross-modal information and provide more potential support for clinical diagnosis. However, effectively preserving both shallow and deep features of each modality while integrating multi-source representations remains challenging, often leading to blurred edges and loss of fine details in fused images. To address this issue, this paper proposes a novel Dual-stream Decoupled Feature parallel alignment fusion network (DDFNet). Built upon an autoencoder (AE) framework, the encoder adopts a dual-branch multi-scale design, consisting of a Detail Feature Extraction Module (DFEM) and a Global Context Feature Extraction Module (GCFEM). These two parallel branches leverage the complementary advantages of CNNs and Transformers to capture local details and global contextual dependencies, respectively, while operating independently without feature interference. In addition, an Improved Multi-scale Cross-alignment and Spatial Attention Fusion module (IMSC-SAF) is introduced to enhance feature integration. It performs complementary feature alignment and cross-attention to strengthen multiscale interactions, while a spatial attention mechanism highlights spatially salient regions. The decoder then reconstructs the fused representation and generates the final fused image. Extensive experiments on publicly available datasets from Harvard Medical School demonstrate that the proposed method achieves competitive fusion performance. In particular, DDFNet consistently obtains the highest MI, VIF, and SSIM values across the evaluated datasets while maintaining competitive results in the remaining objective metrics. Qualitative comparisons further demonstrate that the proposed method effectively preserves complementary anatomical and functional information, producing visually balanced fusion results.

Keywords: decoupled feature extraction; deep learning; CNN and transformer; multimodal medical image fusion network; parallel alignment fusion



Academic Editor: Cosimo Nardi

Received: 11 June 2026

Revised: 7 July 2026

Accepted: 15 July 2026

Published: 20 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Over the past decade or so, the tremendous development in the medical and health field has spurred the production of various imaging sensors, and medical images acquired using different imaging modes are widely used in clinical diagnostic medical applications. Relying on a single imaging modality often fails to provide comprehensive information about human organs or lesions, which is necessary to ensure the accuracy of clinical scenarios and the thoroughness of examinations [1]. At the same time, different imaging

technologies differ in terms of resolution and contrast, and each imaging technology has its limitations. Different imaging modalities provide different information about the same organ in the body [2]. Among them, MRI, X-ray, CT, and US can all show the location, size, and shape of lesions, as well as the morphological and structural changes they cause in surrounding tissues, thus providing more comprehensive anatomical and structural information. To gain a deeper understanding of the biological processes of tumors and their soft tissue and functional information, PET, functional MRI, and SPECT are becoming increasingly common. In the diagnosis and treatment of the same human organ, multimodal medical image fusion technology has been explored as a useful tool for integrating complementary anatomical and functional information from different imaging modalities, thereby supporting clinical scenarios, treatment planning, and disease monitoring [3,4].

Over the past few decades, image fusion methods have evolved into various technical directions [5]. In this process, based on the core technologies they employ, they can be classified into two main types: one is fusion methods based on traditional image processing techniques [6], and the other is fusion methods based on deep learning techniques [7]. The purpose of both is to fuse multiple images of different modalities into an image with comprehensive information. Traditional medical image fusion methods primarily include: spatial-domain methods [8], frequency-domain methods [9], sparse representation methods [10], and saliency-based methods [11], among others. Most traditional image fusion methods are designed manually, relying on prior human understanding of image features to design fusion rules. Traditional methods are limited by their inherent framework, manually designed fusion rules, and complex manual design, which not only restrict performance optimization but also lead to low inference efficiency, making them unsuitable for strict real-time requirements in medical scenarios.

Deep learning-based methods can be divided into end-to-end deep learning methods and non-end-to-end deep learning methods [12,13]. These methods have inherent deep feature learning capabilities, and their end-to-end architecture avoids the manual design of fusion strategies, significantly improving fusion performance and efficiency. Although deep learning-based medical image fusion methods, especially those using independent encoders and multi-scale feature extraction strategies, have made significant progress, there is still a neglected bottleneck in mainstream designs regarding the key issue of how to most effectively preserve and integrate features of different properties.

As summarized in Figure 1, early autoencoder-based multimodal medical image fusion (MMIF) frameworks mainly relied on pixel-level fusion or single-encoder architectures, as shown in Figure 1a,b. More recent methods generally adopt one of two representative feature extraction paradigms. The first directly propagates shallow feature representations into deeper layers for hierarchical encoding (Figure 1c). Although this strategy facilitates progressive feature abstraction, shallow detail features and deep contextual representations become increasingly entangled during hierarchical encoding, leading to the degradation of fine-grained spatial information and weakening the preservation of modality-specific characteristics. The second paradigm extracts shallow and deep features in parallel and subsequently fuses them through direct element-wise addition within a shared feature space (Figure 1d). While this design improves feature aggregation efficiency, it implicitly assumes semantic compatibility between heterogeneous feature representations. Consequently, structural details and contextual information may interfere with each other before fusion, preventing complementary features from being preserved independently and ultimately limiting multimodal feature representation and fusion performance. On the other hand, convolutional neural networks (CNNs) focus excessively on local information, and their limited receptive field makes it difficult to extract global information from an image. While

the Transformer model excels at global modeling by extracting global information, it neglects local details and has higher computational complexity.

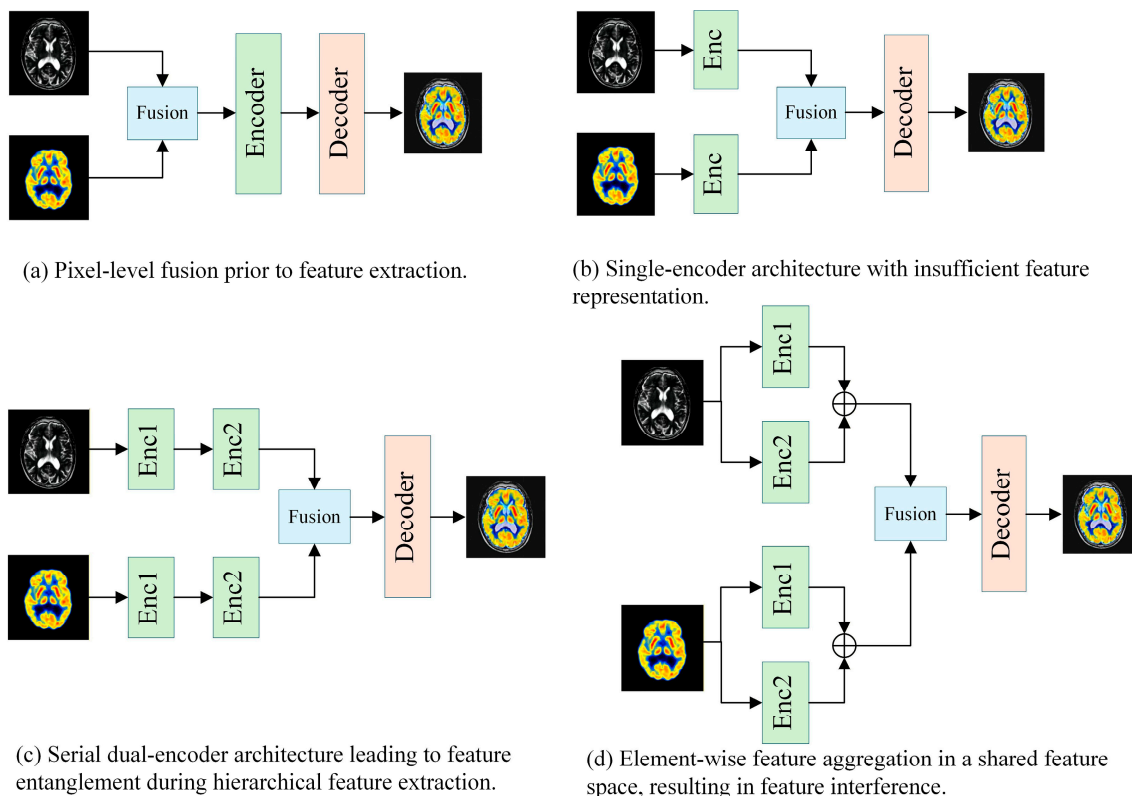


Figure 1. Representative feature extraction paradigms and their inherent limitations in existing autoencoder-based multimodal medical image fusion (MMIF) frameworks.

To address these limitations, we propose a Dual-stream Decoupled Feature Parallel Alignment Fusion Network (DDFNet), in which shallow detail features and deep contextual features are extracted independently to avoid feature entanglement during representation learning. Instead of directly mixing multi-level features in a shared feature space, DDFNet introduces an alignment-guided fusion strategy to enable complementary interaction while mitigating feature interference, thereby improving multimodal medical image fusion performance. The encoder part is a dual-branch, multi-scale design that combines the global context awareness of the Transformer with the local feature extraction efficiency of the CNN. The core of DFEM is lightweight convolution and residual connections, which enables efficient extraction of high-frequency detail information. GCFEM consists of a cascaded hierarchical convolutional encoder (HCE) and a global context transformer (GCT). The HCE progressively encodes hierarchical local representations and performs spatial down-sampling, thereby significantly reducing the computational complexity of the subsequent GCT. Based on the compressed representation, the GCT further captures global context dependencies, thereby extracting deep feature representations. The features are fed in parallel into the IMSC-SAF fusion module, which employs feature alignment and different attention mechanisms to adaptively integrate information from these two different scales, thereby facilitating deep interaction between cross-modal features. The reconstruction module performs precise, high-fidelity decoding of the features output by the fusion module and maps them back to the image pixel space. The novelty of DDFNet does not stem from introducing a new CNN, Transformer, or attention mechanism individually, but from the proposed decoupled feature representation framework, alignment-guided fusion strategy, and the collaborative interaction among these components.

The main contributions of this study are summarized as follows:

- (1) We propose a Dual-stream Decoupled Feature Parallel Alignment Fusion Network (DDFNet) for multimodal medical image fusion. Unlike conventional autoencoder-based fusion frameworks that directly mix shallow and deep representations, DDFNet explicitly decouples detail and contextual feature extraction into two parallel processing streams and performs alignment-guided feature interaction, thereby alleviating feature entanglement and improving feature complementarity.
- (2) We develop a decoupled feature representation strategy that independently models shallow spatial details and deep contextual information. By preserving these complementary representations before fusion, the proposed strategy reduces information interference caused by serial feature extraction or direct feature mixing in a shared feature space.
- (3) We design an alignment-guided fusion mechanism by integrating feature alignment, multi-scale cross-attention, and lightweight spatial attention into a unified fusion module. The proposed interaction strategy enables more effective complementary fusion between decoupled features while mitigating feature mismatch and information interference.

2. Related Work

This section provides a brief review of relevant methods and representative works on deep learning-based multimodal medical image fusion, and outlines the development process of the Transformer architecture in visual tasks.

2.1. Deep Learning-Based Methods for Multimodal Medical Image Fusion

A large number of deep learning-based techniques have been developed in the field of medical image fusion, among which end-to-end fusion methods have become a major research focus. In general, the deep learning-based image fusion pipeline consists of three key stages: feature extraction and processing, feature fusion, and feature reconstruction [12]. Existing deep learning-based multimodal image fusion methods involve a variety of network architectures, mainly including autoencoders (AEs), generative adversarial networks (GANs), convolutional neural networks (CNNs), Transformers, and diffusion models [14]. Since it is extremely difficult to obtain ground-truth images for medical image fusion tasks, most mainstream deep learning-based methods adopt unsupervised training strategies [12]. These methods typically construct loss functions by measuring the differences between the fused image and each source image, thereby driving the optimization of network parameters.

2.2. CNN-Based MMIF Methods

CNNs are among the most widely adopted deep learning models for MMIF due to their strong capability in extracting local spatial features and preserving fine structural details [15]. Liu et al. [16] first introduced CNNs into the image fusion field and proposed a deep convolutional neural network for multi-focus image fusion. Subsequently, Zhang et al. [17] developed IFCNN, a fully convolutional framework that performs feature extraction, feature fusion, and image reconstruction in an end-to-end manner. Xu and Ma et al. [18] proposed EMFusion, an unsupervised medical image fusion network that enhances information preservation through both shallow- and deep-level supervision. Han et al. [19] proposed IE-CFRN, which introduces shallow feature interaction at each stage of hierarchical feature extraction.

However, due to the inherently local receptive field of convolution, CNN-based methods are limited in capturing long-range contextual dependencies, making it difficult to fully model global semantic information in MMIF tasks.

2.3. Transformer-Based MMIF Methods

Transformer, originally proposed by Vaswani et al. [20], captures long-range dependencies through self-attention and has been widely adopted in multimodal medical image fusion following the introduction of the Vision Transformer (ViT) [21] and the Swin Transformer [22], owing to its superior capability for global contextual modeling [22–27]. In MMIF, Tang et al. [28] proposed the Multi-scale Adaptive Transformer (MATR), which enhances global semantic representation through an adaptive Transformer architecture. Ma et al. [29] introduced SwinFusion, a general image fusion framework that combines cross-domain long-range dependency modeling with the Swin Transformer for multimodal image fusion. Wang et al. [30] further proposed SwinFuse, which exploits hierarchical self-attention to improve the interaction between local and global features, thereby enhancing the visual quality of fused images.

Such architectures prioritize global contextual representation, often at the expense of fine-grained structural detail preservation and the independent modeling of modality-specific complementary information.

2.4. Hybrid CNN–Transformer Autoencoder Frameworks

Hybrid CNN–Transformer autoencoder frameworks have emerged as a representative paradigm for multimodal medical image fusion, where feature extraction, feature fusion, and image reconstruction are jointly optimized in an end-to-end manner. Early autoencoder-based MMIF frameworks typically relied on a single encoder for multimodal feature extraction, followed by feature fusion and image reconstruction through a decoder. Such designs, predominantly based on CNNs [17,18] and more recently extended to Transformers [28,30], encode heterogeneous modalities within a unified feature extraction pathway. As a result, modality-specific characteristics tend to become coupled during representation learning, making it difficult to preserve complementary information independently.

To address this limitation, recent studies have increasingly integrated CNNs and Transformers within autoencoder frameworks to jointly exploit local detail perception and global contextual modeling. Representative methods, such as SwinFusion [29], LRFNet [31], and FATFusion [32], have demonstrated improved fusion performance by leveraging the complementary strengths of convolution and self-attention mechanisms.

Nevertheless, most existing hybrid architectures still rely on a unified feature extraction and representation paradigm, which limits the independent preservation of shallow detail features and deep contextual information. Consequently, modality-specific complementary characteristics cannot be fully disentangled, which restricts the expressive capability of learned representations, and most fusion strategies lack an explicit mechanism to address inter-modal feature misalignment, ultimately limiting fusion performance. Therefore, designing an effective feature representation framework that explicitly decouples and collaboratively exploits these complementary characteristics remains an important research direction in multimodal medical image fusion.

We propose a parallel decoupled alignment-fusion paradigm that extracts local and global features via independent dual streams and explicitly rectifies inter-modal spatial misalignment prior to fusion.

3. Proposed Method

To address the aforementioned challenges in MMIF, we propose a fusion network named DDFNet, whose overall architecture is illustrated in Figure 2a. Built upon an AE-based architecture, the proposed network mainly consists of a Detail Feature Extraction Module (DFEM), a Global Context Feature Extraction Module (GCFEM), an advanced improved multiscale cross-alignment and spatial attention fusion module (IMSC-SAF), and an image reconstruction module.

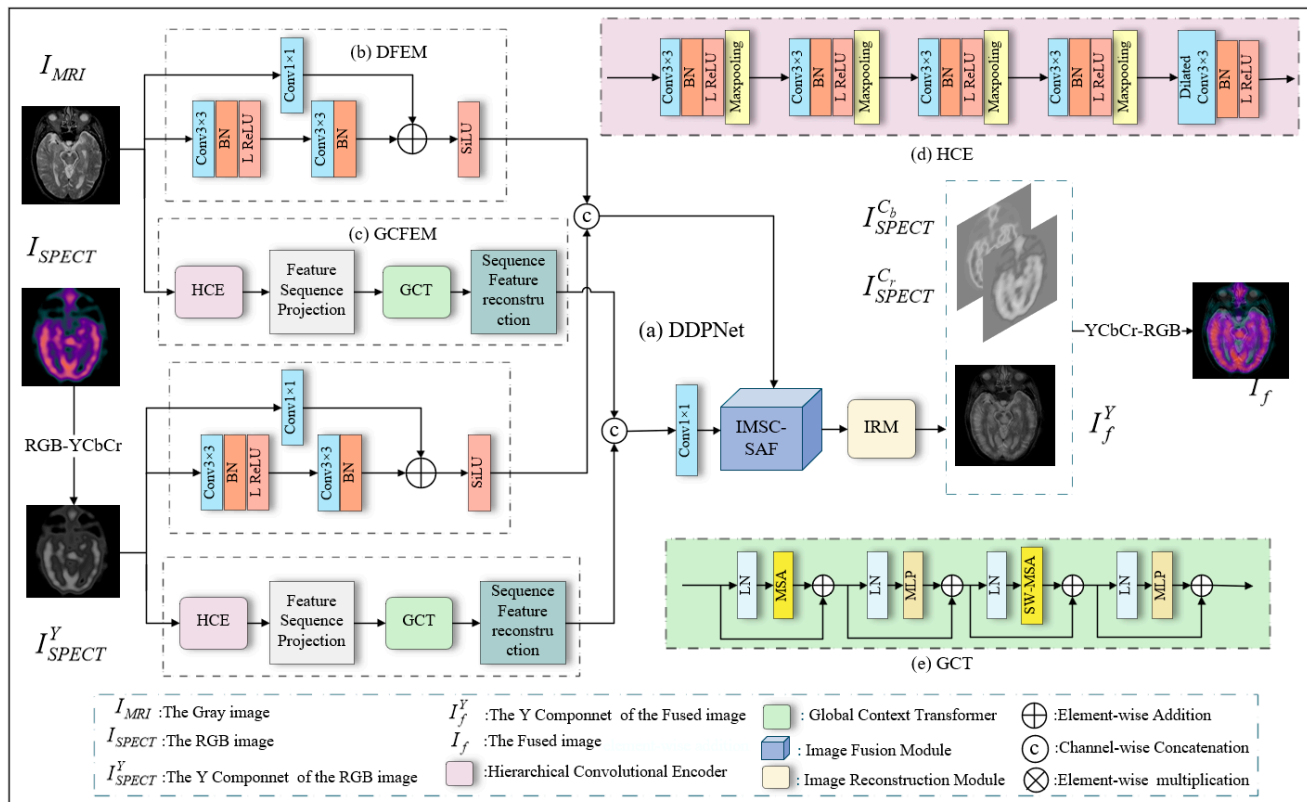


Figure 2. (a) The overall network structure of DDFNet, consisting of DFEM, GCFEM, IMSC-SAF, and IRM; (b,c) The framework of DFEM, GCFEM; (d,e) The framework of HCE, GCT.

RGB-to-YCbCr transformation is a widely used strategy for addressing the fusion problem between three-channel RGB functional images and single-channel grayscale structural images [33,34]. Specifically, let the MRI image be denoted as $I_a \in \mathbb{R}^{H \times W \times C_a}$ and the PET/SPECT image as $I_b \in \mathbb{R}^{H \times W \times C_b}$, where W and H denote the width and height of the image, respectively, and C_a and C_b represent the number of channels. After applying the RGB-to-YCbCr transformation, the PET/SPECT image is decomposed into three channels, denoted as I_b^Y , $I_b^{C_b}$ and $I_b^{C_r}$, respectively. Subsequently, I_a and I_b^Y are fed into the proposed DDFNet. The shallow features extracted from the two inputs by the Detail Feature Extraction Module are concatenated, while the deep features extracted by the Global Context Feature Extraction Module are also concatenated, resulting in unified shallow and deep feature representations. These two aligned feature streams are then treated as explicit and complementary inputs to the subsequent feature fusion and reconstruction modules. The fused image in the Y channel is represented as I_f^Y , and the final RGB fused image I_f is generated through the inverse YCbCr-to-RGB transformation. Therefore, the proposed network performs fusion only on the luminance (Y) channel, while the chrominance components are preserved without participating in the learning process. This design enables the network to

focus on learning complementary structural and functional information in the luminance domain while preserving the original functional color representation.

3.1. Feature Extraction Encoder Module

Feature extraction is a crucial step for the model to understand image information. To enhance the feature saliency of structurally informative regions, a dual-branch multi-scale feature extraction module is proposed to perform specialized processing of feature information at different levels. By employing a parallel-branch structure, the module is capable of simultaneously capturing local detail features and global contextual information in medical images.

Specifically, we design a Detail Feature Extraction Module (DFEM) to preserve spatial details and local structural information during fusion. As shown in Figure 2b, the DFEM employs a lightweight convolutional architecture with residual connection to balance detail representation and computational efficiency. It contains two cascaded convolutional layers followed by BN, where Leaky ReLU and SiLU are used as the activation functions. Both layers use a kernel size of 3×3 and a stride of 1. A 1×1 convolution is further introduced for channel matching in the residual path, which facilitates feature reuse and alleviates gradient vanishing. The base convolution channel number of DFEM is set to 16, and the output feature dimension is maintained at 16 for subsequent feature extraction and fusion. The above operations can be formulated as follows:

$$F_{\text{Out1}} = \text{LReLU}(\text{BN}(\text{Conv}(I_{\text{In}}))) \quad (1)$$

$$F_{\text{Out2}} = \text{BN}(\text{Conv}(F_{\text{Out1}})) \quad (2)$$

$$F_{\text{Out}}^{\text{DFEM}} = \text{SiLU}(F_{\text{Out2}} + \text{Conv}_{1 \times 1}(I_{\text{In}})) \quad (3)$$

where F_{Out1} and F_{Out2} represent the outputs of the first and second convolutional layers, respectively; $F_{\text{Out}}^{\text{DFEM}}$ denotes the output feature generated by the DFEM; LReLU and SiLU denote the activation functions; BN denotes the normalization operation; Conv and $\text{Conv}_{1 \times 1}$ represent the 3×3 and 1×1 convolutions operations, respectively; and I_{In} denotes the input image feature; and “+” denotes the residual connection.

As shown in Figure 2c, the Global Context Feature Extraction Module (GCFEM) consists of a Hierarchical Convolutional Encoder (HCE) and a Global Contextual Transformer (GCT). The HCE employs convolutional layers to progressively extract local contextual representations while reducing feature dimensions, thereby generating compact feature embeddings for subsequent Transformer processing and significantly reducing the computational burden of the self-attention mechanism. Subsequently, the GCT leverages self-attention to capture global contextual relationships and long-range dependencies, thereby improving the global discriminative capability of the features. In this way, the convolutional encoder and the Transformer complement each other, with the former providing efficient local feature modeling and compact representations, and the latter enhancing global contextual reasoning through self-attention.

In the HCE module, a hierarchical progressive structure is adopted to balance receptive field expansion and feature representation capability. As shown in Figure 2d, the network consists of four cascaded “convolution-normalization-activation-pooling” units. In each unit, a 3×3 convolution is used to extract local features and expand the channel dimension, while max pooling performs downsampling. After four downsampling stages, a dilated convolutional layer is introduced to enlarge the receptive field without changing the spatial resolution, thereby capturing richer multi-scale contextual information. The convolutional layers use a kernel size of 3×3 with a stride of 1, the max-pooling layers use a kernel size of 2×2 with a stride of 2, and the dilated convolution uses a kernel size of 3×3 , a padding

of 2, and a dilation rate of 2. The intermediate feature channels are progressively increased from 16 to 32, 64, and 128 to enlarge the receptive field and enhance hierarchical contextual representation. The above process can be formulated as follows:

$$F_{\text{Out}}^{\text{HCE}} = \text{LReLU}(\text{BN}(\text{DilatedConv}(\text{MaxPool}_{2 \times 2}(\text{LReLU}(\text{BN}(\text{Conv}(\text{MaxPool}_{2 \times 2}(\text{LReLU}(\text{BN}(\text{Conv}(\text{MaxPool}_{2 \times 2}(\text{LReLU}(\text{BN}(\text{Conv}(\text{MaxPool}_{2 \times 2}(\text{LReLU}(\text{BN}(\text{Conv}(I_{\text{In}}))))))))))))))))))))) \quad (4)$$

where $F_{\text{Out}}^{\text{HCE}}$ denotes the output feature of the HCE, $\text{MaxPool}_{2 \times 2}$ denotes the 2×2 max-pooling operation, and Conv denotes the 3×3 convolution operation.

Inspired by the patch embedding strategy in the Vision Transformer, we design a Feature Sequence Projection Block (FSPB) to transform the feature maps produced by the HCE into a sequence format compatible with the Transformer architecture. This process involves flattening the spatial dimensions, projecting the features to the target dimension via a linear layer, and adding positional encoding to retain spatial information.

In the GCT module, Transformer blocks derived from the Vision Transformer and Swin Transformer are directly introduced. The specific structure is shown in Figure 2e. The module consists of two Transformer Layer Blocks (TLBs), resulting in a lightweight design. The first TLB contains two residual units: the first is composed of layer normalization (LN) and multi-head self-attention (MSA), while the second consists of LN and a multilayer perceptron (MLP). The second TLB shares the same structure but replaces MSA with shifted window-based multi-head self-attention (SW-MSA), thereby reducing computational complexity relative to global attention. The embedding dimension of the GCT is set to 64, with four attention heads and an MLP ratio of 4.0, corresponding to a hidden dimension of 256. The window size was empirically set to 8×8 , following the common configuration adopted in Swin Transformer-based vision models. To improve computational efficiency, the Transformer operates only on compressed 16×16 feature maps, rather than the original high-resolution features.

$$F_{\text{Out1}}^{\text{TLB}} = \text{MLP}(\text{LN}(\text{MSA}(\text{LN}(F_{\text{Out1}}^{\text{FSP}})) + F_{\text{Out1}}^{\text{FSP}})) + \text{MSA}(\text{LN}(F_{\text{Out1}}^{\text{FSP}})) + F_{\text{Out1}}^{\text{FSP}} \quad (5)$$

$$F_{\text{Out}}^{\text{GCT}} = \text{MLP}(\text{LN}(\text{SWMSA}(\text{LN}(F_{\text{Out1}}^{\text{TLB}})) + F_{\text{Out1}}^{\text{TLB}})) + \text{MSA}(\text{LN}(F_{\text{Out1}}^{\text{TLB}})) + F_{\text{Out1}}^{\text{TLB}} \quad (6)$$

where $F_{\text{Out1}}^{\text{TLB}}$ and $F_{\text{Out}}^{\text{GCT}}$ represent the output of the first TLB and the output of the GCT module, respectively; LN denotes Layer Normalization; MSA denotes the multi-head self-attention operation, MLP denotes the computation process of the multilayer perceptron, and SWMSA denotes the shifted window-based multi-head self-attention operation; $F_{\text{Out}}^{\text{FSP}}$ denotes the output of the FSPB module and “+” denotes the residual connection.

After the GCT, we design a Sequence-to-Feature Map Reconstruction module to restore the Transformer output sequence into a feature map representation. This process is implemented through dimension transposition and reconstruction, which recovers spatial positional information and provides a standard input format for subsequent feature fusion.

3.2. Feature Fusion Module

The fusion module is designed to integrate the features extracted in the previous stage, enabling complementary and collaborative interactions across modalities and feature levels. To this end, we propose an Improved Multiscale Cross-alignment and Spatial

Attention Fusion module (IMSC-SAF) to effectively integrate shallow detail features and deep contextual information. Specifically, the proposed IMSC-SAF consists of a Multiscale Cross-alignment Fusion (MSC-AF) module for feature alignment and cross-scale interaction, followed by a lightweight spatial attention module for spatial refinement of the fused features. The overall architecture is shown in Figure 3.

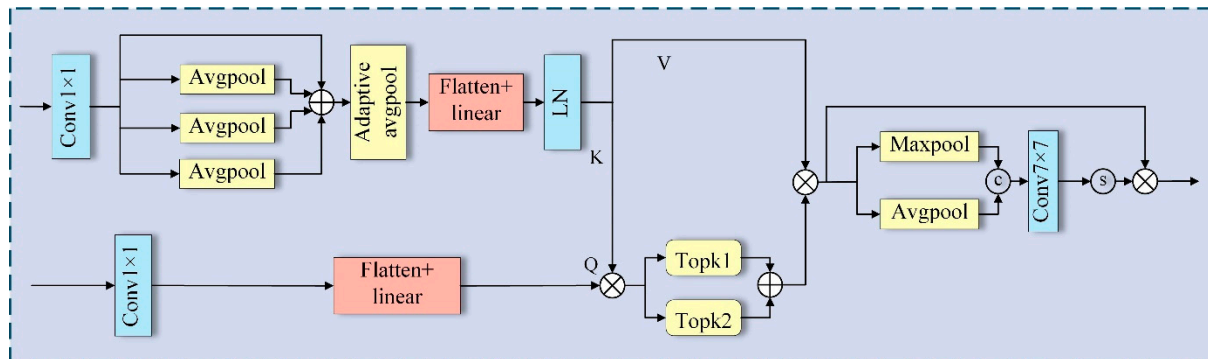


Figure 3. The specific network structure of IMSC-SAF.

Compared with the original MSC module, the proposed MSC-AF introduces three key improvements to enhance feature alignment and cross-modal feature fusion. Based on the original MSC module [35], we first use a 1×1 convolution layer to project the input features from the two branches into a unified 128-channel dimension, and employ adaptive average pooling to align their spatial sizes, thereby addressing the fusion bottleneck caused by feature mismatch. A residual connection is further introduced into the multi-scale pooling branch to preserve the original feature information. Meanwhile, adaptive average pooling not only achieves spatial alignment but also reduces the sequence length required for the subsequent cross-attention mechanism. The improved module is termed Multiscale Cross-attention with Alignment Fusion (MSC-AF). While retaining the advantages of multi-scale perception and sparse cross-attention fusion in the original module, MSC-AF enables more efficient feature alignment and preliminary fusion. These three modifications distinguish MSC-AF from the original MSC module while retaining its advantages in multi-scale perception and sparse cross-attention fusion, thereby enabling more effective feature alignment and preliminary cross-modal feature fusion.

To further refine the fused representations and enhance their focus on key anatomical structures, a lightweight spatial attention module is cascaded after the MSC-AF module. This module aggregates cross-channel information from the preliminarily fused feature maps to generate a spatial attention map, which adaptively emphasizes informative anatomical regions while suppressing less relevant responses. The specific operation is formulated as follows:

$$M_S(F_{Out}^{MSC-AF}) = \text{Mul}(\text{Sig}(\text{Conv}_{7 \times 7}(\text{CAT}(\text{AvgPool}(F_{Out}^{MSC-AF}), \text{MaxPool}(F_{Out}^{MSC-AF})))), F_{Out}^{MSC-AF}) \quad (7)$$

where F_{Out}^{MSC-AF} denotes the output of the MSC-AF module, CAT denotes the channel-wise concatenation operation, Mul denotes the multiplication operation, Sig denotes the Sigmoid activation function, $\text{Conv}_{7 \times 7}$ denotes the 7×7 convolution operation, and AvgPool and MaxPool denote the average pooling and max pooling operations, respectively.

3.3. Feature Reconstruction Decoder Module

The image reconstruction module is the final step in AE-based multimodal image fusion. To restore the spatial size of the fused image, bilinear interpolation is adopted for

upsampling. Each upsampling block consists of an upsampling operation followed by a convolutional layer, where the convolution is used to refine the up sampled features, recover image details and clarity, and reduce the feature dimension. In this way, smooth upsampling is achieved while artifacts are alleviated. The detailed structure is shown in Figure 4.

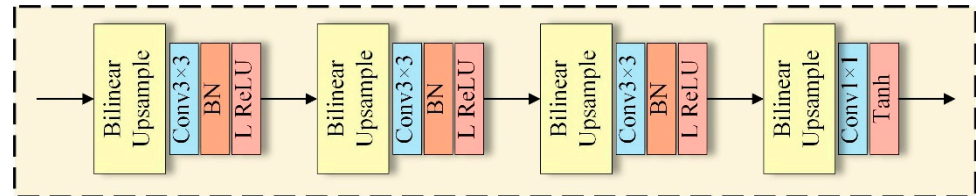


Figure 4. The framework of image reconstruction module (IRM).

3.4. Loss Function

To supervise the training and preserve the original image features, a composite loss function consisting of structural similarity (SSIM) loss, content loss, and gradient loss is employed. The structural similarity index (SSIM) measures image similarity by comparing luminance, contrast, and structure. Based on the design principles of the SSIM loss function [36], it is defined as follows:

$$L_{SSIM} = 1 - SSIM(I_f^Y, I_a) + \alpha \cdot (1 - SSIM(I_f^Y, I_b^Y)) \quad (8)$$

where I_a and I_b^Y represent the source images, I_f^Y represents the fused image, α is a weighting factor used to balance the contribution of each source image in the SSIM calculation, and $SSIM(\cdot)$ denotes the structural similarity operation. The SSIM is defined as follows:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (9)$$

where μ , σ_{xy} , and σ^2 denote the mean, covariance, and variance, respectively, and C_1 and C_2 are constants used to stabilize the computation.

L_{SSIM} is adopted to constrain the structural similarity between the fused image and the source images. To further preserve fine details, we introduce both content loss and gradient loss. Instead of forcing the fused image to approximate a single-source modality, the pixel-wise maximum operation is adopted as a surrogate supervision to encourage the preservation of the dominant complementary responses from both modalities. Since MRI mainly provides anatomical structures, whereas CT, PET, and SPECT provide complementary density or functional information, the pixel-wise maximum helps reduce the loss of salient modality-specific information during optimization while avoiding excessive bias toward either modality. These two pixel-wise losses jointly form the edge loss function [37]. Specifically, the content loss L_c is used for intensity constraint and is defined as follows:

$$L_c = \frac{1}{H \times W} \|I_f^Y - \max(I_a, I_b^Y)\|_1 \quad (10)$$

where H and W denote the height and width of the fused image I_f^Y , respectively, and $\max(\cdot)$ and $\|\cdot\|_1$ denote the maximum operation and the L_1 norm, respectively.

The gradient loss L_g is employed to preserve texture details and is defined as follows:

$$L_g = \frac{1}{H \times W} \| |\nabla I_f^Y| - \max(|\nabla I_a|, |\nabla I_b^Y|) \|_1, \quad (11)$$

where ∇ represents the Sobel operator for gradient calculation, and $|\cdot|$ denotes the absolute value operation. Accordingly, the edge loss is defined as follows:

$$L_E = L_C + \beta L_g \quad (12)$$

The parameter β controls the trade-off between the content loss and the gradient loss. Accordingly, the overall loss function is defined as follows:

$$L_{\text{total}} = L_{\text{SSIM}} + \gamma L_E \quad (13)$$

where γ is a hyperparameter for balancing the total loss L_{total} so as to obtain high-quality fused images.

4. Experiments and Results

In this section, we first describe the datasets, training details, comparison methods, and evaluation metrics. Then, qualitative and quantitative comparisons are conducted between DDFNet and several state-of-the-art fusion methods. Finally, the experimental results are further analyzed and discussed.

4.1. Datasets and Experimental Settings

The Harvard Medical Image Dataset is a widely used dataset for multimodal medical image fusion [28,29,34]. All the images used in this experiment are from the publicly available Harvard dataset. We downloaded 413 pairs of MRI and SPECT images, and 167 pairs of MRI and CT images. Each image has a resolution of 256×256 pixels. The datasets were split into training and testing sets with a 9:1 ratio, resulting in 372 training and 41 testing pairs for MRI-SPECT, and 150 training and 17 testing pairs for MRI-CT. To avoid data leakage, all datasets were split at the subject level, ensuring that image pairs from the same subject do not appear in both training and testing sets. Additionally, 32 pairs of MRI and PET images were downloaded for generalization experiments, and were not involved in training or parameter tuning. No additional preprocessing was performed on the data, except for converting PET/SPECT images to the YCbCr color space and separating the Y channel.

All experiments in this study were implemented and trained using the PyTorch2.0 framework on an Intel Core i5-13600KF CPU and an NVIDIA GeForce RTX 4070Ti GPU. During the training phase, the Adam optimizer was used with a learning rate of 0.0001, a batch size of 4, and a total of 100 epochs. The hyperparameters α and γ in the loss function were set to 1 and 0.1, respectively. Since the two source modalities are considered equally important during fusion, α was set to 1 to assign equal weights to both source images in the SSIM loss. Li et al. [37] recommended setting β to 10, and we adopted this value accordingly to balance the content and gradient losses. The parameter γ was empirically set to 0.1 based on preliminary experiments. A larger γ places excessive emphasis on the edge loss, which degrades the overall fusion quality and weakens structural consistency.

4.2. Comparison Methods and Evaluation Metrics

To objectively and quantitatively evaluate the fusion results, six mainstream evaluation metrics are employed: Entropy [38] (EN), Spatial Frequency [39] (SF), Normalized Weighted Edge Information [40] ($Q^{AB/F}$), Mutual Information [41] (MI), Visual Information Fidelity [42] (VIF), and Structural Similarity Index [36] (SSIM).

A brief introduction to each metric is as follows: EN measures the amount of information contained in the fused image. SF measures the gradient distribution of the fused image to evaluate the level of detail and texture preservation. $Q^{AB/F}$ measures the amount

of edge information transferred from the source images to the fused image. MI measures the amount of information transferred from the source image to the fused image. VIF quantifies the similarity and mutual information between the source and fused images in the feature space. SSIM is used to model the level of distortion and loss in the image, thereby reflecting the structural similarity between images.

To validate the effectiveness of the proposed method, we compare it with one traditional method and six deep learning-based methods: LRD [43], IFCNN [17], SwinFuse [30], MATR [28], M4Fnet [44], LRFNet [31], and FATFusion [32]. For all comparison methods, we used either publicly available implementations or re-implemented them according to the original papers. All deep learning-based methods were implemented using the official codes and retrained on the same training datasets with the default hyperparameter settings recommended by the original authors. Qualitative and quantitative results are provided for each fusion task, with LRD being the traditional method.

4.3. Analysis of Results

4.3.1. Qualitative Evaluation

We apply the proposed method along with seven existing methods to the fusion tasks of SPECT-MRI, PET-MRI, and CT-MRI, selecting nine image pairs for visualization analysis. Three fusion results are presented for each modality. Additionally, to provide a clearer view of the images, a region of interest from each image is zoomed in for better observation.

Figure 5 presents the fusion results of three pairs of MRI and SPECT images using eight fusion methods. The fused images generated by the traditional method LRD exhibit significant artifacts. Both LRD and IFCNN show noticeable color shifts compared with the source functional images. SwinFuse preserves the chromatic appearance of the functional images but loses some anatomical structural details. M4FNet improves structural feature extraction compared with SwinFuse but does not adequately preserve fine edge details from the MRI images. FATFusion changes the intensity distribution of the MRI regions. The fusion results of MATR and LRFNet show relatively limited distortions, although slight artifacts remain. In comparison, the proposed method preserves more distinguishable anatomical structures while maintaining a more consistent intensity distribution. The fused images also retain the functional characteristics of the source SPECT images without introducing obvious visual artifacts.

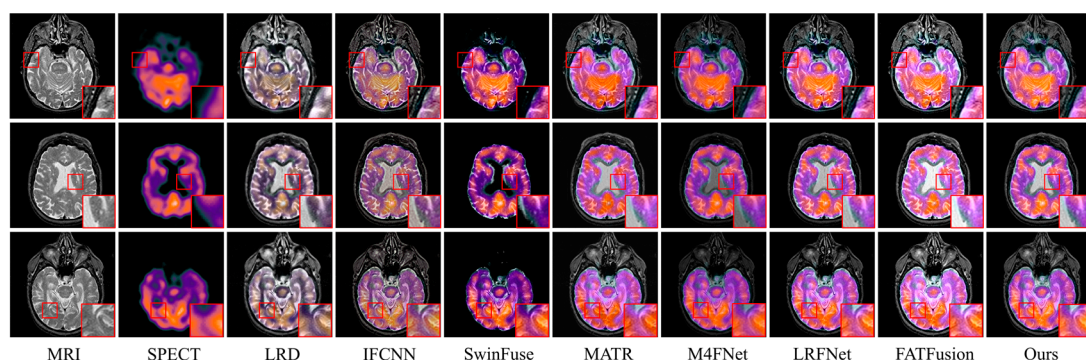


Figure 5. Qualitative comparison of our DDFNet with seven methods on three MRI and SPECT image pairs. The red boxes indicate the regions enlarged for detailed comparison.

To further evaluate the generalization capability of the proposed method, experiments were conducted on MRI–PET image pairs. As shown in Figure 6, LRD still exhibits considerable artifacts, while both LRD and IFCNN suffer from noticeable color distortion. SwinFuse shows limited capability in preserving anatomical structural information. M4FNet does not adequately preserve fine texture details in the MRI images and produces relatively low-

intensity edge regions. MATR effectively preserves anatomical information but weakens the intensity characteristics of the PET images. FATFusion changes the intensity distribution of the MRI regions. The proposed method effectively preserves anatomical texture details while maintaining the functional characteristics presented in the PET images, resulting in visually consistent fusion results with well-preserved edge structures.

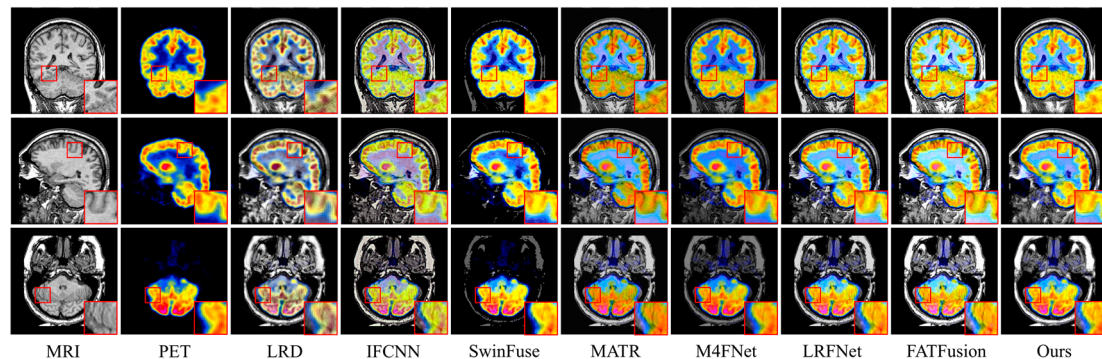


Figure 6. Qualitative comparison of our DDFNet with seven methods on three MRI and PET image pairs. The red boxes indicate the regions enlarged for detailed comparison.

Figure 7 shows the fusion results of three pairs of MRI and CT images using eight fusion methods. LRD suffers from insufficient image clarity, resulting in detail loss and visible artifacts. Although the remaining deep learning-based methods generally produce satisfactory fusion results, differences remain in preserving edge details and smooth tissue regions. IFCNN produces relatively dark fused images, while SwinFuse overemphasizes CT information and weakens important MRI structural details. M4FNet loses part of the MRI structural information and reduces the visibility of dense CT structures. MATR, LRFNet, and FATFusion weaken the intensity characteristics of CT images, leading to slight structural degradation. In comparison, the proposed method better preserves anatomical edge structures while maintaining the complementary information from both MRI and CT images, producing visually consistent fusion results with fewer observable artifacts.

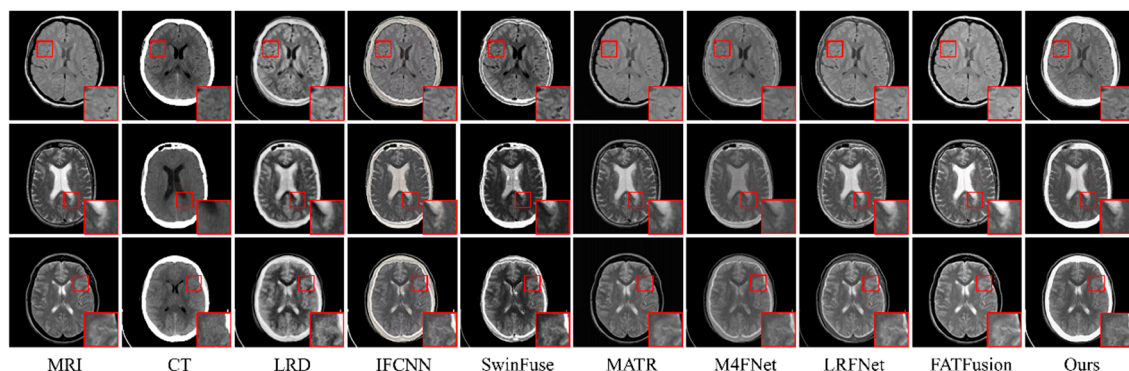


Figure 7. Qualitative comparison of our DDFNet with seven methods on three MRI and CT image pairs. The red boxes indicate the regions enlarged for detailed comparison.

4.3.2. Quantitative Evaluation

This paper conducts a quantitative analysis of eight fusion methods using six evaluation metrics, where higher values indicate better performance for EN, SF, $Q^{AB/F}$, MI, VIF, and SSIM.

Tables 1–3 present the objective evaluation results of the comparison experiments on the MRI-SPECT, MRI-PET, and MRI-CT datasets. The traditional method LRD achieves

the highest EN across all three datasets. The traditional method LRD achieves the highest EN across all three datasets. Since EN mainly measures information richness rather than information quality, higher EN values may also reflect the preservation of additional redundant or noisy information. Specifically, the proposed method maintains the highest values of MI, VIF and SSIM across all three datasets. The superior MI values indicate that the proposed network is capable of effectively exploiting complementary information among different modalities. The improvement in VIF further demonstrates that the fused images can better preserve perceptually important information that is sensitive to the human visual system. Meanwhile, the highest SSIM values further verify the superiority of the proposed method in maintaining overall anatomical structures and global tissue consistency. In contrast, the SF and $Q^{AB/F}$ metrics mainly emphasize the preservation of high-frequency textures, gradient variations, and edge information. DDFNet focuses more on deep contextual information modeling and cross-modal global feature fusion, without explicitly introducing edge enhancement or gradient sharpening mechanisms. Although the proposed method is slightly inferior to some edge-oriented approaches in terms of local high-frequency edge-related metrics, it emphasizes balanced integration of anatomical structures and functional information through hierarchical feature extraction and adaptive feature fusion. Consequently, the proposed method achieves competitive performance in terms of overall structural consistency, functional continuity, and visual naturalness.

Table 1. The quantitative comparison results between the proposed DDFNet method and seven competitors in SPECT and MRI image fusion; Bold indicates the best value and “Ours” denotes the proposed DDFNet method.

Methods	EN	SF	$Q^{AB/F}$	MI	VIF	SSIM
LRD	5.5469	21.8757	0.7742	2.6996	0.2901	0.4303
IFCNN	4.8473	22.2159	0.3396	2.8989	0.3131	0.7264
SwinFuse	4.1252	16.1417	0.3683	1.7471	0.1512	0.6811
MATR	5.0662	18.3919	0.5710	2.0142	0.3543	0.3219
M4FNet	4.7532	11.6992	0.4499	1.9846	0.2397	0.7288
LRFNet	4.6630	19.5461	0.7924	1.6743	0.2975	0.7308
FATFusion	5.0428	24.5715	0.6298	3.0511	0.9169	0.7042
Ours	5.0047	18.6447	0.4458	3.1610	0.9385	0.7481

Table 2. The quantitative comparison results between the proposed DDFNet method and seven competitors in PET and MRI image fusion; bold indicates the best value.

Methods	EN	SF	$Q^{AB/F}$	MI	VIF	SSIM
LRD	6.1949	32.9063	0.7570	2.4181	0.2825	0.4358
IFCNN	5.3622	35.8321	0.3581	2.7374	0.2873	0.7084
SwinFuse	3.8945	23.4588	0.2254	1.6154	0.1652	0.6443
MATR	5.6839	27.1778	0.5967	1.7983	0.3187	0.3505
M4FNet	5.2278	18.7558	0.4961	2.2216	0.3346	0.6607
LRFNet	5.2656	30.4594	0.7807	1.7109	0.3109	0.6918
FATFusion	5.5687	37.3221	0.6397	2.7614	0.6937	0.6911
Ours	5.5065	29.1240	0.4495	2.8490	0.8541	0.7123

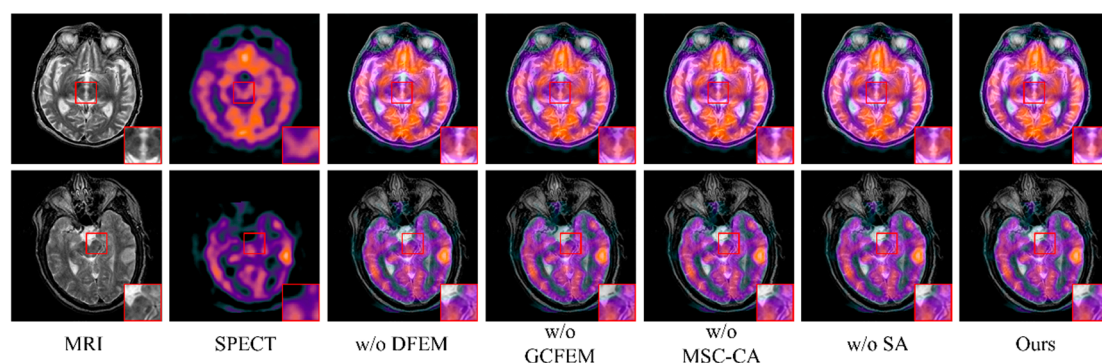
Overall, DDFNet demonstrates competitive overall performance across the three datasets. In particular, it consistently achieves the highest MI, VIF, and SSIM values while maintaining competitive performance on the remaining evaluation metrics, objectively demonstrating the effectiveness and robustness of the proposed method. The model demonstrates promising cross-modal transfer, warranting further validation on larger PET cohorts.

Table 3. The quantitative comparison results between the proposed DDFNet method and seven competitors in CT and MRI image fusion; bold indicates the best value.

Methods	EN	SF	$Q^{AB/F}$	MI	VIF	SSIM
LRD	5.2818	28.6095	0.5651	2.6891	0.2247	0.4706
IFCNN	4.5850	33.5509	0.2436	2.8040	0.2392	0.7332
SwinFuse	4.1550	36.7577	0.4677	1.9260	0.1690	0.7130
MATR	4.8523	18.3892	0.5419	2.0255	0.3629	0.2817
M4FNet	4.2225	19.1717	0.4002	1.8746	0.1851	0.7286
LRFNet	4.4852	22.4555	0.4492	1.5167	0.2312	0.7320
FATFusion	4.5514	26.0713	0.5415	2.8637	0.6662	0.7157
Ours	4.5798	24.9870	0.3272	2.9158	0.7899	0.7468

4.4. Ablation Experiment

To validate the rationality of different modules in DDFNet, ablation experiments were designed focusing on the two stages of feature extraction and feature fusion. The evaluation was conducted using the seven metrics mentioned earlier. In addition, the visual effects of the ablation experiments on the MRI-PET modality are shown in Figure 8. Specifically, the experimental setups are as shown below.

**Figure 8.** The fusion results of ablation experiments in MRI-SPECT. The red boxes indicate the regions enlarged for detailed comparison.

Without using DFEM, the network relies solely on GCFEM and the subsequent fusion modules for feature modeling and fusion. Without using GCFEM, only the shallow features extracted by DFEM are used for fusion. With the basic feature alignment operation retained, the multi-scale sparse cross-attention mechanism (MSC-CA) is removed. The spatial attention mechanism in the IMSC-SAF module is removed, leaving only the multi-scale cross-fusion structure, thus achieving the complete DDFNet network model.

Table 4 presents the results of the ablation experiments, demonstrating that the removal of key modules leads to a performance degradation in fusion, validating the necessity of each module in DDFNet. Specifically, the absence of the feature extraction module weakens the model's ability to model detail information or global context, while removing the attention mechanism in the fusion stage limits the effective integration of cross-modal information and modeling of key regions. In contrast, the complete DDFNet model shows more stable and balanced performance across various evaluation metrics, indicating that the proposed modular design effectively enhances the quality of multimodal medical image fusion. The fused images exhibit a good balance in terms of detail clarity, structural consistency, and modality complementarity, confirming the rationality and effectiveness of the proposed network structure design.

Table 4. Results of ablation experiments in MRI-SPECT; bold indicates the best value.

Modality	Experiments	EN	SF	$Q^{AB/F}$	MI	VIF	SSIM
MRI-SPECT	w/o DFEM	4.9908	18.6674	0.4318	3.1533	0.9187	0.7394
	w/o GCFEM	4.9841	18.1141	0.4305	3.0590	0.8241	0.7255
	w/o MSC-CA	4.9778	18.7225	0.4374	3.1479	0.9284	0.7283
	w/o SA	4.9966	18.7064	0.4383	3.1501	0.9279	0.7365
	Full DDFNet (Ours)	5.0047	18.6447	0.4458	3.1610	0.9385	0.7481

5. Conclusions

In this paper, we propose a novel network called DDFNet for multimodal medical image fusion. The network is designed to address the entanglement of detail information and global contextual information during feature transmission. Unlike conventional architectures that rely on a unified feature extraction paradigm, the proposed method introduces a dual-stream decoupled feature learning strategy to explicitly separate the modeling of local detail information and global contextual dependencies.

Specifically, the Detail Feature Extraction Module (DFEM) is designed to preserve fine-grained spatial details and local structural information, while the Global Context Feature Extraction Module (GCFEM) is constructed with a hierarchical convolutional encoder and a Transformer-based global context modeling branch to capture long-range dependencies efficiently. These two complementary feature streams are then jointly fed into the IMSC-SAF fusion module, where cross-attention and spatial attention mechanisms are employed to achieve adaptive feature alignment and multi-scale information integration. Extensive experimental results and ablation studies further validate the effectiveness of each component. In particular, DFEM significantly enhances fine-grained detail preservation, while GCFEM improves global contextual modeling capability. Moreover, the proposed IMSC-SAF module effectively facilitates cross-modal feature alignment and adaptive multi-scale fusion, leading to superior fusion performance.

The experimental results demonstrate that the proposed architecture achieves improved fusion performance in both qualitative and quantitative evaluations, particularly in terms of detail preservation and contextual consistency.

To further evaluate the computational efficiency of the proposed method, we report its model complexity. The proposed DDFNet contains 1.32 M parameters and achieves 3.60 GFLOPs with an inference time of 14.7 ms per image, demonstrating that the incorporation of Transformer-based modules does not significantly increase computational overhead due to the use of compressed feature representations. The proposed network demonstrates competitive performance on various medical image fusion tasks and achieves a reasonable trade-off among model size, efficiency, and fusion quality. Nevertheless, some limitations still remain. In future work, we will develop a more general fusion network for multimodal imaging scenarios in different domains, with greater emphasis on the practical requirements of downstream vision tasks or a radiologist-led reader study to provide expert subjective assessment and further establish the clinical value of the proposed fusion approach.

Author Contributions: Conceptualization, P.H.; methodology, P.H.; software, P.H.; validation, P.H. and M.C.; data curation, P.H.; writing—original draft preparation, P.H.; writing—review and editing, M.C. and B.L.; supervision, C.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in [<http://www.med.harvard.edu/AANLIB/home.html>], accessed on 10 June 2026]. The implementation code, trained model weights, and train/test split information will be made available upon reasonable request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Haidekker, M.A. X-ray projection imaging. In *Medical Imaging Technology*; Springer: New York, NY, USA, 2013; pp. 13–35.
- El-Gamal, F.E.Z.A.; Elmogy, M.; Atwan, A. Current trends in medical image registration and fusion. *Egypt. Inform. J.* **2016**, *17*, 99–124. [[CrossRef](#)]
- Rajalingam, B.; Priya, R.; Scholar, R. Review of multimodality medical image fusion using combined transform techniques for clinical application. *Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud.* **2018**, *7*, 1–8.
- Kavita, P.; Alli, D.R.; Rao, A.B. Study of image fusion optimization techniques for medical applications. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 136–143. [[CrossRef](#)]
- Yin, M.; Liu, X.; Liu, Y.; Chen, X. Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampling shearlet transform domain. *IEEE Trans. Instrum. Meas.* **2019**, *68*, 49–64. [[CrossRef](#)]
- Yang, Y.; Que, Y.; Huang, S.; Lin, P. Multimodal sensor medical image fusion based on type-2 fuzzy logic in NSCT domain. *IEEE Sens. J.* **2016**, *16*, 3735–3745. [[CrossRef](#)]
- Liu, R.; Liu, Y.; Wang, H.; Li, J.; Hu, K. SAFFusion: A saliency-aware frequency fusion network for multimodal medical image fusion. *Biomed. Opt. Express* **2025**, *16*, 2459–2481. [[CrossRef](#)] [[PubMed](#)]
- Chen, J.; Li, X.; Luo, L.; Mei, X.; Ma, J. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Inf. Sci.* **2020**, *508*, 64–78. [[CrossRef](#)]
- Georgieva, V.M.; Petrov, P.P.; Tsvetkova, D.S.; Laskov, L.B. MRI/SPECT image fusion of brain based on multi-scale wavelet decomposition. In Proceedings of the 56th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), Sozopol, Bulgaria, 16–18 June 2021; pp. 85–88. [[CrossRef](#)]
- Wang, J.; Ren, P.; Yang, K.; Qin, C.; Zhang, X. Image fusion based on gradient regularized convolution sparse representation. In Proceedings of the 9th Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 23–26 September 2018; pp. 1–4.
- Veshki, F.G.; Ouzir, N.; Vorobyov, S.A. Image fusion using joint sparse representations and coupled dictionary learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 8344–8348. [[CrossRef](#)]
- Zhou, T.; Cheng, Q.; Lu, H.; Li, Q.; Zhang, X.; Qiu, S. Deep learning methods for medical image fusion: A review. *Comput. Biol. Med.* **2023**, *160*, 106959. [[CrossRef](#)] [[PubMed](#)]
- Chen, W.; Sun, K.; Li, Q.; Xie, K.; Ni, X. Research progress of multimodal medical image fusion methods. *Chin. J. Radiol. Health* **2023**, *32*, 580–585.
- Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
- Mehta, S.; Singh, A. Hybrid deep learning for multi-modal image fusion: Elevating brain tumor detection. In Proceedings of the 2nd International Conference on Recent Trends in Microelectronics, Automation, Computing and Communication Systems (ICMACC), Prayagraj, India, 19–21 December 2024; pp. 453–457. [[CrossRef](#)]
- Liu, Y.; Chen, X.; Peng, H.; Wang, Z. Multi-focus image fusion with a deep convolutional neural network. *Inf. Fusion* **2017**, *36*, 191–207. [[CrossRef](#)]
- Zhang, Y.; Liu, Y.; Sun, P.; Yan, H.; Zhao, X.; Zhang, L. IFCNN: A general image fusion framework based on convolutional neural network. *Inf. Fusion* **2020**, *54*, 99–118. [[CrossRef](#)]
- Xu, H.; Ma, J. EMFusion: An unsupervised enhanced medical image fusion network. *Inf. Fusion* **2021**, *76*, 177–186. [[CrossRef](#)]
- Han, Y.; Nie, R.; Cao, J.; Liang, S.; Pan, L. IE-CFRN: Information exchange-based collaborative feature representation network for multi-modal medical image fusion. *Biomed. Signal Process. Control* **2023**, *86*, 105301. [[CrossRef](#)]
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 10012–10022.

23. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers and distillation through attention. In Proceedings of the 38th International Conference on Machine Learning (ICML), Virtual Event, 18–24 July 2021; pp. 10347–10357.
24. Wang, W.; Xie, E.; Li, X.; Fan, D.P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 568–578.
25. Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P.H.S. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
26. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021.
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
28. Tang, W.; He, F.; Liu, Y.; Duan, Y. MATR: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Trans. Image Process.* **2022**, *31*, 5134–5149. [[CrossRef](#)] [[PubMed](#)]
29. Ma, J.; Tang, L.; Fan, F.; Huang, J.; Mei, X.; Ma, Y. SwinFusion: Cross-domain long-range learning for general image fusion via Swin transformer. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1200–1217. [[CrossRef](#)]
30. Wang, Z.; Chen, Y.; Shao, W.; Li, H.; Zhang, L. SwinFuse: A residual Swin transformer fusion network for infrared and visible images. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 5003812. [[CrossRef](#)]
31. He, D.; Li, W.; Wang, G.; Huang, Y.; Liu, S. LRFNet: A real-time medical image fusion method guided by detail information. *Comput. Biol. Med.* **2024**, *173*, 108381. [[CrossRef](#)] [[PubMed](#)]
32. Tang, W.; He, F. FATFusion: A functional–anatomical transformer for medical image fusion. *Inf. Process. Manag.* **2024**, *61*, 103687. [[CrossRef](#)]
33. Tang, W.; Wang, L.; Liu, Y. Green fluorescent protein and phase contrast image fusion via dual attention residual network. In Proceedings of the IEEE International Conference on Medical Imaging Physics and Engineering (ICMIPE), Hefei, China, 12–14 November 2021; pp. 1–6. [[CrossRef](#)]
34. Xu, Y.; Wang, Z.; Wu, S.; Zhan, X. MPCT: A medical image fusion method based on multi-scale pyramid convolution and transformer. *Biomed. Signal Process. Control* **2025**, *101*, 107129. [[CrossRef](#)]
35. Ma, J.; Jiang, W.; Tang, X.; Zhang, X.; Liu, F.; Jiao, L. Multiscale sparse cross-attention network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5605416. [[CrossRef](#)]
36. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
37. Li, J.; Liu, J.; Zhou, S.; Zhang, Q.; Kasabov, N.K. GeSeNet: A general semantic-guided network with couple mask ensemble for medical image fusion. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 16248–16261. [[CrossRef](#)] [[PubMed](#)]
38. Roberts, J.W.; Van Aardt, J.A.; Ahmed, F.B. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *J. Appl. Remote Sens.* **2008**, *2*, 023522. [[CrossRef](#)]
39. Eskicioglu, A.M.; Fisher, P.S. Image quality measures and their performance. *IEEE Trans. Commun.* **1995**, *43*, 2959–2965. [[CrossRef](#)]
40. Petrovic, V.; Xydeas, C. Objective image fusion performance characterisation. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV), Beijing, China, 17–20 October 2005; pp. 1866–1871.
41. Qu, G.; Zhang, D.; Yan, P. Information measure for performance of image fusion. *Electron. Lett.* **2002**, *38*, 313–315. [[CrossRef](#)]
42. Han, Y.; Cai, Y.; Cao, Y.; Xu, X. A new image fusion performance metric based on visual information fidelity. *Inf. Fusion* **2013**, *14*, 127–135. [[CrossRef](#)]
43. Li, X.; Guo, X.; Han, P.; Wang, X.; Li, H.; Luo, T. Laplacian redecomposition for multimodal medical image fusion. *IEEE Trans. Instrum. Meas.* **2020**, *69*, 6880–6890. [[CrossRef](#)]
44. Ding, Z.; Li, H.; Guo, Y.; Zhou, D.; Liu, Y.; Xie, S. M4FsNet: Multimodal medical image fusion network via multi-receptive-field and multi-scale feature integration. *Comput. Biol. Med.* **2023**, *159*, 106923. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.