DCATNet: Dilated Convolution Attention Transformer Network for Medical Image Fusion

Zenghui Wang, Mingliang Gao*, Lina Liu

School of Electrical and Electronic Engineering,
Shandong University of Technology,
Zibo, China,
E-mail: sdut_zenghuiwang@163.com
mlgao@sdut.edu.cn
liulina@sdut.edu.cn
*Corresponding author

Xiangqin Zeng

Zibo Central Hospital,
Zibo, China
E-mail: lucky_fairy365@163.com

Qilei Li

School of Electronic Engineering and Computer Science,
Queen Mary University of London,
London, United Kingdom
E-mail: q.li@qmul.ac.uk

Monire Norouzi

Computer Technology Program,
Vocational School,
Haliç University,
Istanbul, Turkey
E-mail: monirenorouzi@halic.edu.tr

Abstract: Medical image fusion is dedicated to extracting structural and functional information from medical images. However, existing medical image fusion methods usually rely on convolutional operations and ignore long-distance information transmission. To address this problem, we propose a Dilated Convolutional Attention Transformer Network (DCATNet) for medical image fusion. Specifically, to enhance the long-term dependence of the network on the input image, a transformer (TF) module is built. At the same time, a Dilated Convolutional Channel Attention (DCCA) module is built to realize the accurate extraction of feature and multi-scale information. This module combines the CPA module with the expansion convolution to enhance the robustness of the model. This enables the proposed method to handle the complexities of long-distance information transfer without losing important contextual and structural details. Experimental results demonstrate that the DCATNet outperforms competitors

and proves its potential in medical image fusion for long-distance information transfer processes. Meanwhile, the results highlight the potential of DCATNet to advance medical image fusion, and it can lead to better clinical outcomes and more accurate diagnoses.

Keywords: Medical image fusion; Transformer; Multi-scale features; Dilated convolution; Long-distance information transmission.

1 Introduction

Multimodal image fusion stands as a prominent research focus within the realm of computer vision. This field integrates information from diverse modalities into a unified representation, which can obtain more comprehensive and richer information [20]. Thanks to the significant advantages of capturing human body information and medical planning, multimodal medical image fusion has emerged as a pivotal and consequential topic in image fusion.

By the inherent characteristics of imaging mechanisms, multimodal medical images focus on structural information and functional information [3]. Structural information within medical images offers details about anatomical structures. For instance, Computed Tomography (CT) excels at delineating diseased areas within organs, and Magnetic Resonance Imaging (MRI) furnishes information on soft tissues. Besides, functional information within medical images provides metabolic information about organs. For example, the Positron Emission Tomography (PET) is instrumental in tumor localization. Employing multimodal image fusion to integrate information from multimodal medical images can improve the accuracy of diagnosis of patient conditions, detect biomarkers, and promote the development of personalized medicine.

The advancement of medical image fusion has rapidly accelerated in response to society's growing demands for clinical medicine and multimodal medical imaging. Medical image fusion can be categorized into traditional and deep learning methods [36] [17]. Traditional methods typically employ fundamental image preprocessing techniques for initial manipulation. Subsequently, pivotal data is systematically gathered, and fusion algorithms are employed to amalgamate the individual images into a consolidated image. Nevertheless, traditional methods may exhibit suboptimal performance in addressing intricate image fusion tasks [28]. These methods often struggle with the complexity of weight map generation through fusion rules.

In recent years, the potent capabilities of deep learning in feature extraction and data analysis have garnered considerable attention [29]. Numerous deep learning-based methods have emerged [9]. Ram *et al.* [16] employed an unsupervised deep fusion algorithm for multi-exposure medical images. Liu *et al.* [11] employed a multi-focus image fusion method to establish a direct mapping from the input source image to the focus image. Additionally, Hou *et al.* [8] fused low-frequency coefficients through a convolutional neural network, which processes medical images from CT and MRI. Nevertheless, these methods primarily apply to multi-exposure image fusion during the fusion process, increasing computational costs.

Although deep learning-based methods have demonstrated success in image fusion tasks, certain methods fail to effectively represent the long-term dependencies of source images. This results in insufficient extraction of global context information. Besides, many approaches tend to concentrate on specific scales and disregard others, which leads to reduced resolution in medical images and the loss of multi-scale information. These approaches need to effectively handle the extraction and fusion of information across multiple scales and long distances. To address the aforementioned challenges, we proposed the Dilated Convolution Attention Transformer Network (DCATNet). A TF module is adopted to extract global contextual information which enhances the model's ability to capture long-distance dependencies within the input images. Subsequently, a Dilated Convolution Channel Attention (DCCA) module is built to extract multi-scale and local channel information. This module utilizes dilated convolutions to gather information across various scales and then address the issue of scale variance that often plagues existing methods.The DCATNet can provide higher-quality, more detailed, and comprehensive images, which improves diagnostic accuracy, optimizes treatment planning, and enhances precision in image-guided interventions. The contributions of this work are summarized as follows:

1. We propose an end-to-end network named DCATNet. It leverages local, global, and multi-scale information to achieve feature complementarity between multimodal images.

2. We propose a DCCA module to capture local features and mult-scale features. It utilizes dilated convolutions to distill multi-scale information and uses channel pooling attention to capture local feature information.

3. We propose a Transformer module to augment the model's perception and grasp of global information. It facilitates the rearrangement and adjustment of global features to improve the overall structural capture of the image.

4. Extensive experiments demonstrate that the proposed method outperforms state-of-the-art competitors. The proposed method can better capture comprehensive local and global features.

The rest of the paper is organized as follows. Section 2 reviews the relevant work on traditional medical image fusion, deep learning-based medical image fusion, and medical diagnosis-based methods. Section 3 presents the proposed model in detail. The experimental results and analysis are described in Section 4. The conclusion and future work are presented in Section 5.

2 Related Work

2.1 Traditional Medical Image Fusion

Traditional medical image fusion methods focus on processing various types of information in the source images, *e.g.,* texture, gradient, detail, and color. The decomposition strategies and fusion rules are usually designed based on different fusion requirements. Dinh *et al.* [4] divided the image into low-frequency parts and high-frequency parts. These parts were processed using local energy functions and the Chameleon swarm algorithm, respectively.

Faragallah *et al.* [5] employed principal component analysis, singular value decomposition, and image fusion to build an efficient fusion network. Gao *et al.* [6] introduced a saliency detection-based fusion model. They utilized the Graph-Based Visual Saliency algorithm for saliency calculation and adopted the particle swarm algorithm to optimize the function of the fuzzy logic system.

Although traditional image fusion methods succeed in specific scenarios, the feature fusion strategies of these methods may limit the fusion performance [33]. The generation of weight maps through fusion rules is a complex process that involves intricate calculations. This poses computational challenges and difficulties in parameter settings for traditional methods [23].

2.2 Deep Learning-based Medical Image Fusion

Deep learning is used as a focused methodology in a variety of fields [1]. It has become an important tool in modern medical imaging due to its ability to learn automatically. Convolutional Neural Network (CNN) has been widely adopted in multi-modal image fusion. Wang *et al.* [21] introduced a medical image fusion method that can apply CNN to image fusion directly. Xia *et al.* [22] introduced a multimodal medical image fusion method by employing a deep convolutional approach. They decomposed the input medical images into images with different frequencies, and the multiple-frequency images were merged to obtain a fused medical image. Zhao *et al.* [35] introduced a medical image fusion network to solve the problems of losing wanted high-frequency input information. Guo *et al.* [7] introduced a conceptual medical image fusion framework encompassing most supervised image fusion analysis methods.

Generative Adversarial Networks (GANs) have also been applied to medical image fusion. Based on GANs, Nai *et al.* [15] utilizes a triple convolutional neural network to fuse medical images. This network continuously trains the discriminator to generate image features, and the generator and discriminator are trained against each other to generate fused images. Zhang *et al.* [32] introduced an adversarial network for medical image fusion. The network uses adaptive decision blocks to give the generated fusion results the same distribution as the source image. Wang *et al.* [19] proposed a cyclic consistency model, termed DiCyc, that enables signal quality to be maintained when synthesized data is aligned with source data. Zhai *et al.* [30] proposed a GAN network that combines self-attention with multi-scale to fuse the CT and MRI images.

Although the aforementioned methods are effective, they fail to represent long-term dependencies, which lead to reduced resolution and loss of multi-scale information. To address the above issues, we constructed a fusion network named DACTNet to improve feature extraction capabilities that handle multiple scales and long-distance dependencies.

3 Methodology

3.1 Overview

The architecture of the DCATNet is shown in Fig. 1. We utilize a 5×5 convolution layer to extract shallow features. Then, a dual attention residual (DAR) module [18] is adopted to extract local feature information. To extract global context information, two TF modules are designed for global context information extraction. After that, we adopt

another DAR module to access high-level semantic features. To obtain an image with multi-scale information and local feature information, we propose a DCCA module. It can utilize complementary and multi-scale information in images efficiently. After processing information in all four channels, we integrate the information from each pathway. At the end of this network, we employ a 1×1 convolution layer to reduce the dimensionality and generate the fusion result.

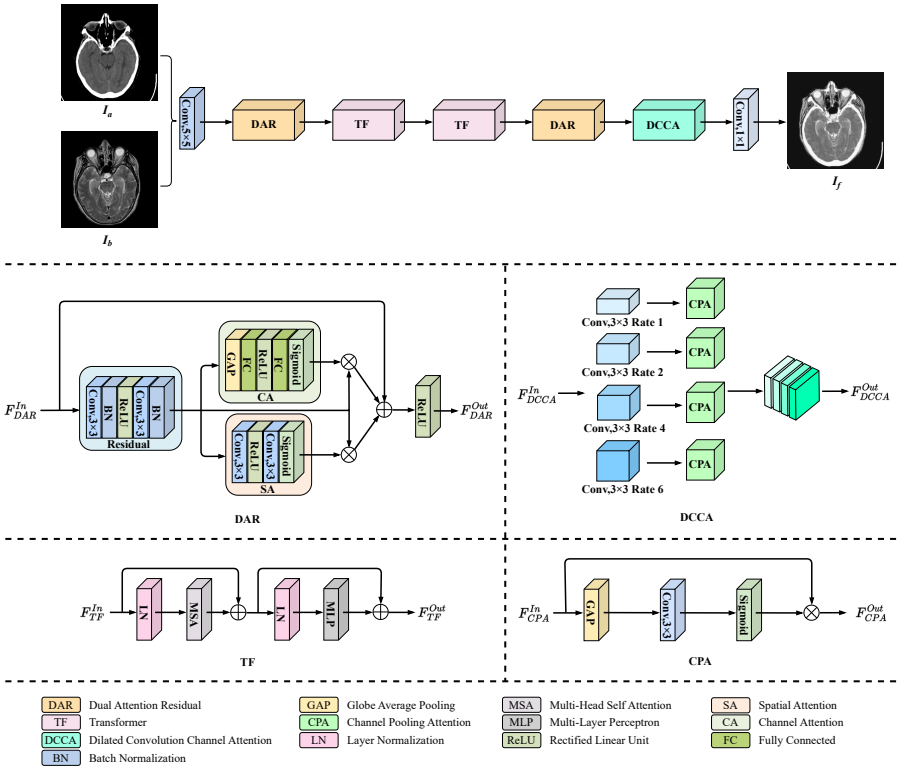


Figure 1 Architecture of the proposed DCATNet for medical image fusion.

3.2 Dual Attention Residual

As shown in Fig. 1, the DAR consists of three parts, namely residual, Channel Attention (CA), and Spatial Attention (SA). The input information is passed through the convolution layer and Batch Normalization (BN) layer to extract the texture information when it enters the residual block. This process can be formalized as,

$$F_R^{Out} = BN\left(Conv_{3\times3}^{\delta,\delta}\left(ReLU\left(BN\left(Conv_{3\times3}^{\delta,\delta}\left(F_{DAR}^{In}\right)\right)\right)\right)\right), \tag{1}$$

where F_{DAR}^{In} is the input of the DAR, and F_R^{Out} is the output of the residual block. $Conv_{3\times3}^{\delta,\delta}$ is a 3×3 convolution which input and output channel is δ. $ReLU(\cdot)$ is the rectified linear unit, and $BN(\cdot)$ is the batch normalization. With the output of the residual block, the F_R^{Out} is processed in parallel by CA and SA.

The CA block is adopted to improve the representational power of DCATNet. The input of the CA, F_{CA}^{In}, passes through a Global Average Pooling (GAP) layer to generate channel descriptors and enters a Fully Connected (FC) layer for dimensionality reduction. Then, the dimensionality is increased through a ReLU and another FC layer. Finally, the output of the CA, F_{CA}^{Out}, is obtained by sigmoid activation. The SA block is used to extract spatial attention. This process is formalized as,

$$F_{SA}^{Out} = Sigmoid\left(Conv_{3\times3}^{\delta/R,1}\left(ReLU\left(Conv_{3\times3}^{\delta,\delta/R}\left(F_{R}^{Out}\right)\right)\right)\right), \tag{2}$$

where F_{SA}^{Out} is the output of the SA. $Sigmoid(\cdot)$ is the Sigmoid activation operation. $Conv_{3\times3}^{\delta/R,1}$ is 3×3 convolution, in which the input channel is δ/R and output channel is 1. The $Conv_{3\times3}^{\delta,\delta/R}$ is 3×3 convolution in which the input channel is δ and output channel is δ/R. Finally, the F_{DAR}^{Out} is obtained by integrating the output information. This process can be formalized as,

$$F_{DAR}^{Out} = ReLU(F_{R}^{Out}\odot F_{CA}^{Out} + F_{R}^{Out}\odot F_{SA}^{Out} + F_{DAR}^{In}), \tag{3}$$

where the F_{DAR}^{Out} is the output of the DAR. The $\odot$ is the multiplication operation.

3.3 Dilated Convolution Channel Attention

The DCCA module is built to capture local features and mult-scale features. It adopts dilated convolutions to extract multi-scale information and uses Channel Pooling Attention (CPA) to capture local feature information. The DCCA module contains two parts: dilated convolution lays and CPA. The dilated convolution is adopted to extract multi-scale information, and CPA is employed to extract local feature information. Compared with traditional convolution operations, different expanded convolution kernels can increase the perceptual field of view and reduce information loss during the fusion process.Firstly, different dilated convolutions are utilized to extract multi-scale information. The input information is fed into four parallel dilated convolutions with rates r = 1, 2, 4, 6. Different dilation rates have different characteristics. A smaller rate is equivalent to traditional convolution, while a larger rate enables the convolutional kernel to cover wider receptive fields on the input. This process is formulated as,

$$\begin{cases} F_{CPA1}^{In} = Conv_{3\times3,rates=1}(F_{DCCA}^{In}), \\ F_{CPA2}^{In} = Conv_{3\times3,rates=2}(F_{DCCA}^{In}), \\ F_{CPA3}^{In} = Conv_{3\times3,rates=4}(F_{DCCA}^{In}), \\ F_{CPA4}^{In} = Conv_{3\times3,rates=6}(F_{DCCA}^{In}), \end{cases} \tag{4}$$

where F_{CPA}^{In} is the input of the CPA, and F_{DCCA}^{In} is the input of the DCCA.

The CPA module in DCATNet plays a crucial role in capturing channel information and local feature information for medical image fusion. It comprises a global pooling layer, a convolutional layer, and an activation function layer. Firstly, a global pooling operation is applied to the input to extract global information from each channel. Then, a 1×1 convolution operation is performed and a Sigmoid function is adopted to learn the channel attention. This process can be formalized as,

$$F_{x} = Sigmoid(Conv_{3\times3}(\text{AvgPool}(F_{CPAt}^{In}))), \tag{5}$$

where F_x is the result after the activation function operation. The $Conv_{3\times3}$ is the 3×3 convolution. The AvgPool$(\cdot)$ denotes global average pooling. The F_{CPAt}^{In} represents four input parallel pathways of the CPA, and t denotes the t-th input paths. CPA enhances image details and contrast by generating a weight to emphasize information in an important channel. It enhances the visual fidelity of the image by capturing local feature information. Ultimately, it is integrated with multi-scale information to obtain a more comprehensive fusion effect. To capture detailed information about the images, we design a multiplication operation in DCCA. It is formulated as:

$$F_{CPAt}^{Out} = F_x \odot F_{CPAt}^{In}, \tag{6}$$

where the F_{CPAt}^{Out} is four parallel pathways output of the CPA. The t corresponds to 1, 2, 3, and 4.

Then, we integrate four dilated convolution pathways to obtain the output. This process is formulated as,

$$F_{DCCA}^{Out} = Cat([F_{CPA1}^{Out}, F_{CPA2}^{Out}, F_{CPA3}^{Out}, F_{CPA4}^{Out}, U_5]), \tag{7}$$

where F_{DCCA}^{Out} is the output of the F_{DCCA}, and U_5 is the upsampling operation. $Cat(\cdot)$ is the integration of channels.

3.4 Transformer Module

The global context information refers to the global features and background information contained in an image. To enhance the capacity of the model to capture the Global context information, we build the Transformer (TF) module. The TF modules are specifically designed to address the limitation of traditional convolutional operations, which struggle with long-range dependencies. The TF module comprises four layers and two additional operations. To extract global contextual information from images, we incorporate MSA and MLP layers. Furthermore, an LN layer is introduced before the two layers to enhance the stability of the model. The formula for the additive operation associated with MSA is formulated as,

$$F_{MSA} = F_{TF}^{In} + \text{MSA}\left(\text{LN}\left(F_{TF}^{In}\right)\right), \tag{8}$$

where F_{TF}^{In} is the input of the TF module. $LN(\cdot)$ represents layer normalization, and MSA$(\cdot)$ signifies the multi-head self-attention mechanism. The F_{MSA} denotes the output of MSA. The formula for the additive operation connected to MLP is formulated as,

$$F_{TF}^{Out} = F_{MSA} + \text{MLP}\left(\text{LN}\left(F_{MSA}\right)\right), \tag{9}$$

where MLP$(\cdot)$ is the multi-layer perception, and F_{TF}^{Out} is the output of the TF module.

When the input information is transferred to the TF module, it decomposes into non-overlapping $M \times M$ windows based on their size, height, width, and the number of channels. These windows can generate $\frac{HW}{M^2} \cdot M^2 \cdot C$ features, where $\frac{HW}{M^2}$ denotes the number of windows. Then, we used linear transformations to transform the input data dimensionally and obtained Q, K, and V. This process is formulated as,

$$\begin{aligned} Q &= W_F \cdot F_Q, \\ K &= W_F \cdot F_K, \\ V &= W_F \cdot F_V, \end{aligned} \tag{10}$$

where Q, K, and V represent the query, key, and value metrics, and F_Q, F_K, and F_V are projection matrices, correspondingly. W_F is the feature of the window and $F_W \in \mathbb{R}^{M^2 \times C}$. We use the self-attention mechanism to compute attention within the window. This process can be formulated as,

$$\text{Attention}(Q, K, V) = \text{S}\left(QK^T/\sqrt{d} + B\right)V, \tag{11}$$

where B is a learnable relative positional encoding. $S(\cdot)$ is the softmax normalization function. $\sqrt{d}$ is a scaling factor for dimension to control the scaling of attention scores.

3.5 Loss Function

The loss function of the proposed DCATNet is formulated as,

$$\mathcal{L} = \mathcal{L}_{\text{P}} + \alpha \cdot \mathcal{L}_{\text{S}} + \beta \cdot \mathcal{L}_{\text{G}}, \tag{12}$$

where $\mathcal{L}_{\text{P}}$, $\mathcal{L}_{\text{S}}$, and $\mathcal{L}_{\text{G}}$ is the pixel loss, structural loss, and gradient loss, respectively. α and β are the weights.

A pixel loss function is designed to balance the fidelity of generated images. The pixel loss function measures the pixel intensity differences between generated and real images. It is formulated as,

$$\mathcal{L}_{\text{P}} = \frac{1}{HW}\left\|I_f - I_a\right\|_F^2 + \lambda \cdot \frac{1}{HW}\left\|I_f - I_b\right\|_F^2, \tag{13}$$

where $\|\cdot\|_F$ is the matrix Frobenius norm. H and W are the height and width of the source images, respectively. λ is a control parameter.

To quantify the structural similarity, a Structural Similarity Index Measurement (SSIM) loss function is designed. It is defined as,

$$\mathcal{L}_{\text{S}} = 1 - \text{SSIM}\left(I_f, \max\left\{I_a, I_b\right\}\right), \tag{14}$$

where SSIM$(\cdot)$ is a measure of structural similarity, and the $max\{,\}$ represents performing a maximum selection.

A gradient loss function is designed to obtain local variation information in the image. The gradient loss function is formulated as,

$$\mathcal{L}_{\text{G}} = \left\|\nabla I_f - \max\left\{\nabla I_a, \nabla I_b\right\}\right\|_2, \tag{15}$$

where ∇ represents the computation of image gradients. $\|\cdot\|_2$ denotes the matrix l_2-norm.

4 Experiments

4.1 Datasets and Experimental Setup

The training experiment was built on 160 pairs of accurately registered CT and MRI images and 245 pairs of PET and MRI images collected from the Harvard Brain Atlas [1]. Twenty-four pairs of CT-MRI images and PET and MRI images were employed for testing.

The training datasets were cropped into blocks with a size of 64×64, and these image patches were normalized to the range of [0,1]. Also, the cropping of image blocks had a stride of 10. The learning rate was set to $1e-3$. Epochs and batch size were set to 100 and 128, respectively. We used the Adam optimizer for parameter updates. The values of α and β in Eq. (12) are set to 100 and 10, and λ in Eq. (13) is set to 1. The values of δ in Eq. (1) and R in Eq. (2) are set to 16 and 4, respectively. The entire network was implemented using the PyTorch framework and trained on an NVIDIA GeForce RTX 3090 Ti GPU.

We compare the proposed method with other state-of-the-art methods, including CSF [26], CUFD [24], FusionGAN [13], GAN-FM [34], GANMcC [14], RFN-Nest [10], SDNet [31], U2Fusion [25], STDFusionNet [12], MUFusion [2] and DATFuse [18]. We conducted experiments using the open-source code of each compared method and compared their parameters with the proposed approach.

4.2 Evaluation Metrics

Four metrics are adopted to evaluate fusion performance, namely mutual information (MI), standard deviation (SD), visual information fidelity (VIF), and Q_{abf} [27]. MI quantifies the amount of information transmitted from the original image to the fused image. A higher MI indicates higher structural similarity between the fused and source images. SD is used to measure the contrast of the fused image. A higher SD indicates that the image has higher contrast, improving image quality. VIF is an indicator of the visual information fidelity between the fused and source images. A higher VIF signifies the distortion of the image is less during the fusion process. Q_{abf} is an indicator of image fusion that measures the image's ability to retain edges. A higher Qabf indicates stronger edge preservation in the fused image.

4.3 Experiments on CT-MRI Datasets

4.3.1 Quantitative Evaluation

The main goal of this experiment was to evaluate the performance of DCATNet in fusing CT and MRI images. Table 1 displays the quantitative results in four metrics. The results indicate that DCATNet performs best in MI, SD, and VIF. Compared with the DATFuse, the proposed DCATNet improves the MI by 6.7%, SD by 24.5%, and VIF by 6.6%, respectively. For the Q_{abf} metric, the proposed DCATNet ranks in second place. This indicates that our DCATNet is comparable, and it can handle edge detail information well. The CT-MRI result indicates that our DCATNet approach can effectively preserve the structural information in MRI images and functional information in CT images. This result ensures clinicians access comprehensive diagnostic information in a single image.

4.3.2 Qualitative Evaluation

The qualitative fusion results of three representative CT and MRI image pairs are shown in Fig. 2. The first two columns are the CT and MRI source images, while the remaining columns display the fusion results from different methods. The first row shows that the fusion results of other competitors preserve detailed dense structure information in the CT image, but only GAN-FM, DATFuse, SDNet, and the proposed DCATNet well-preserved the edge details of the MRI image. In the second row, only CUFD, GAN-FM, MUFusion,

Table 1 Quantitative comparison results of the proposed DCATNet with other competitors on CT and MRI image fusion. (The best and second-best results are marked in red and blue, respectively.)

Method	MI ↑	SD ↑	VIF ↑	Q_{abf} ↑
CSF [26]	2.732	61.076	0.368	0.345
CUFD [24]	2.846	43.797	0.379	0.277
FusionGAN [13]	2.363	33.787	0.227	0.116
GAN-FM [34]	2.816	79.355	0.368	0.301
STDFusionNet [12]	3.254	55.692	0.479	0.458
GANMcC [14]	2.696	54.686	0.346	0.256
U2Fusion [25]	2.585	55.700	0.337	0.458
RFN-Nest [10]	2.605	63.295	0.330	0.209
SDNet [31]	2.562	48.362	0.357	0.480
MUFusion [2]	2.718	76.812	0.393	0.420
DATFuse [18]	3.249	67.393	0.488	0.513
Ours	3.468	83.909	0.520	0.508

DATFuse, and the DCATNet maintain detailed MRI information. In the third row, the GAN-FM and DATFuse have image distortion with the saturation being too high. Compared to other competitors, the proposed DCATNet can preserve the density structure information of CT images and texture feature information of MRI images. The fused images of DCATNet can retain the dense structural details of CT images and the detailed information of MRI images.

The CT-MRI experiment aimed to compare the effectiveness of different models in preserving critical information from CT and MRI images.

4.4 Experiments on PET-MRI Datasets

4.4.1 Quantitative Evaluation

The quantitative evaluation results of PET and MRI image fusion are presented in Table 2. The results show the DCATNet ranks first in terms of MI, SD, and VIF. It proves that our method effectively preserves higher contrast and similarity in the images. Compared with the DATFuse, the proposed DCATNet improves the MI by 5.5%, SD by 7.4%, and VIF by 32.9%, respectively. Regarding the Q_{abf} metric, DCATNet ranked second place and only 2.1% less than SDNet. This indicates that the proposed method is comparable and can effectively preserve the edge image details.

4.4.2 Qualitative Evaluation

The comparison results for three pairs of images from PET and MRI are shown in Fig. 3. The first two columns are the source images of PET and MRI, followed by the fusion results from different methods. In the first row, the fusion results combine the characteristics of both modalities and preserve color and detail information. However, while retaining color information, the contrast varies across images. Only CUFD, GAN-FM, DATFuse, SDNet, U2Fusion, MUFusion, and DCATNet preserve the texture information from MRI in the fusion process. The second row shows that GAN-FM, DATFuse, SDNet, U2Fusion, and

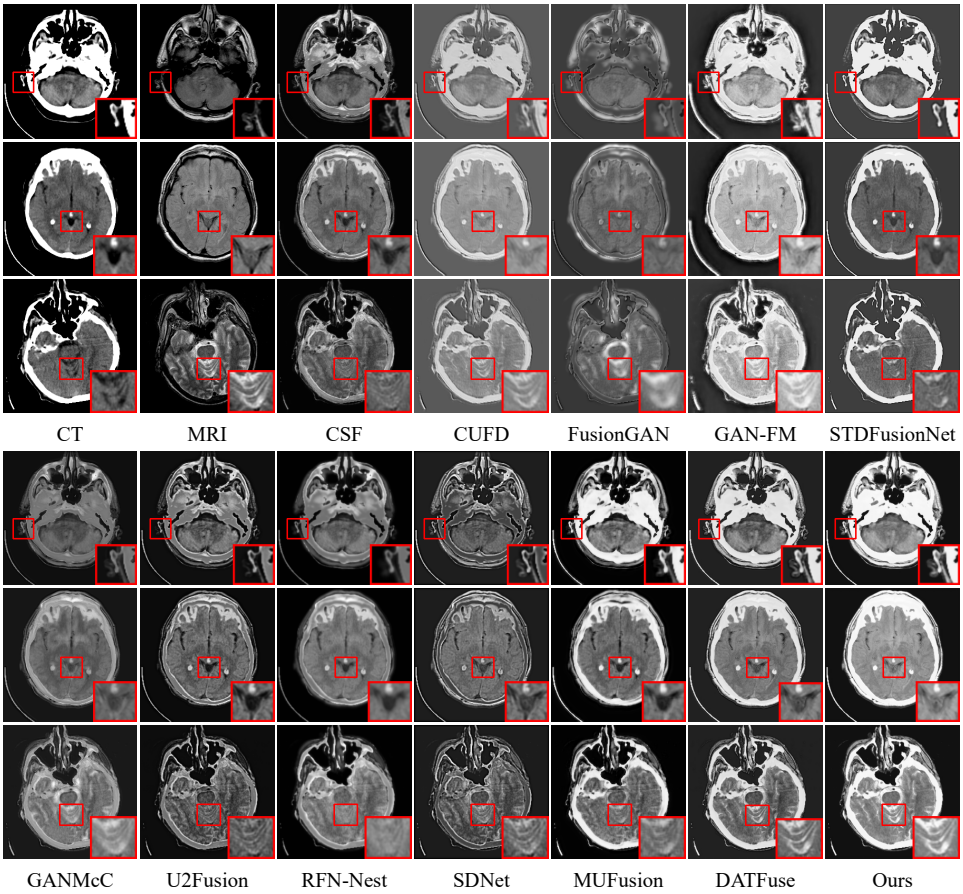


Figure 2 Qualitative evaluation results of the other counterparts on three typical image pairs from the CT and MRI image pairs.

the proposed DCATNet exhibit lower resolution and some color distortion. In the third row, CUFD, GAN-FM, and U2Fusion show a certain degree of attenuation in edge detail information in the fusion results. On the other hand, the DCATNet successfully preserves MRI detail information without introducing color distortion.

The fused images maintained the color information from PET images and the texture details from MRI images without distortion. This combination ensures that critical diagnostic details are preserved and enhanced.

This experiment aimed to test DCATNet's capability to fuse PET and MRI images. The fusion of these modalities aims to improve diagnostic and treatment planning accuracy.

4.5 Efficiency Analysis

To validate the computational complexity of the DCATNet, we performed the comparative experiment on the CT-MRI dataset with a 3080Ti GPU. We adopted two indicators, namely Params and FLOPs, to evaluate the computational complexity of the proposed DCATNet and other competitors.

Table 2 Quantitative comparisons results of the proposed DCATNet with other competitors on PET and MRI image fusion. (The best and second-best results are marked in red and blue, respectively.)

Method	MI ↑	SD ↑	VIF ↑	Q_{abf} ↑
CSF [26]	2.909	72.724	0.530	0.382
CUFD [24]	2.599	53.758	0.488	0.405
FusionGAN [13]	2.231	60.720	0.310	0.155
GAN-FM [34]	2.704	82.874	0.442	0.470
STDFusionNet [12]	2.486	66.166	0.404	0.341
GANMcC [14]	2.822	61.742	0.481	0.302
U2Fusion [25]	2.719	68.386	0.479	0.459
RFN-Nest [10]	2.764	73.524	0.489	0.186
SDNet [31]	2.774	65.621	0.520	0.634
MUFusion [2]	2.392	72.753	0.449	0.432
DATFuse [18]	3.026	88.931	0.419	0.369
Ours	3.191	95.510	0.557	0.621

Table 3 The computational complexity comparisons results of the proposed DCATNet with other competitors. (The best and second-best results are marked in red and blue, respectively.)

Method	Params ↓	FLOPs ↓
CSF [26]	0.0617	37.9728
CUFD [24]	0.9530	0.0019
STDFusionNet [12]	0.2825	0.0006
U2Fusion [25]	0.6592	86.4382
RFN-Nest [10]	4.7915	163.5680
SDNet [31]	0.0671	0.0001
MUFusion [2]	0.5547	51.3147
DATFuse [18]	0.0108	9.5767
Ours	0.0111	9.7864

The comparison results are shown in Table 3. The results indicate that DCATNet performs well in Params and FLOPs. It ranks second place in Params, which is only next to the DATFuse. Meanwhile, it scores 9.7864 in FLOPs, which ranks the fifth place. This means the proposed method can effectively reduce computational complexity while retaining good performance.

4.6 Ablation Studies

To investigate the contributions and necessity of the components in the DCATNet, a series of ablation experiments were built. The ablation experiments are verified in MI, SF, VIF, and Q_{abf} to maintain unified experimental standards. Table 4 presents the quantitative ablation results of the CT and MRI image fusion. “baseline” refers to the vanilla model that only has a framework. "baseline + DAR" reports that add a DAR module to the framework. "baseline + DAR + DAR" means that one more DAR is added to the model. "baseline + DAR + TF + DAR" denotes inserting the TF between DARMs. We added the DAR module to enhance the

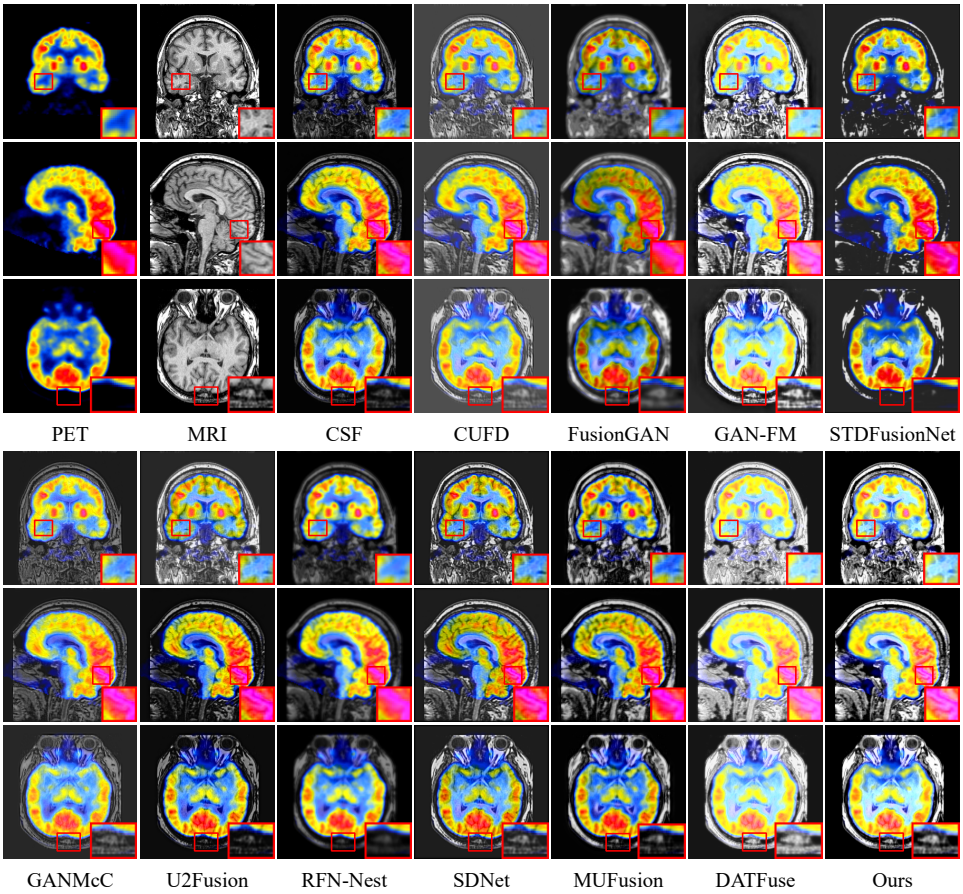


Figure 3 Qualitative evaluation results of the other counterparts on three typical image pairs from the PET and MRI image pairs.

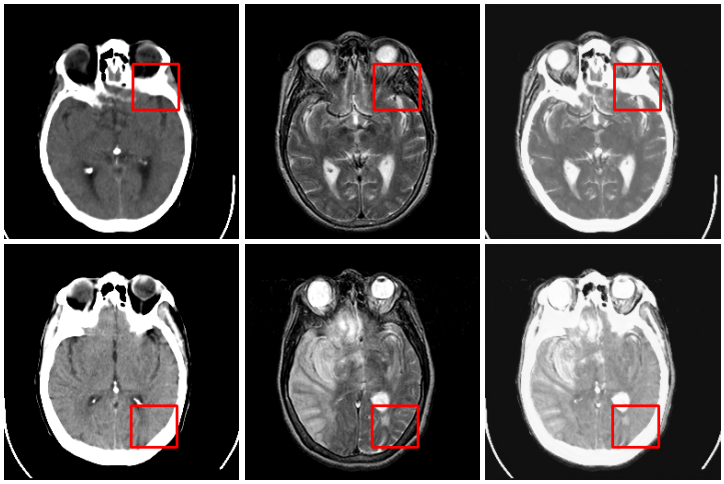
network's feature extraction capability. The model scores 3.346 and 0.509 in MI and VIF, respectively. However, The SD decreased to 66.333. It implies that the contrast information has been affected. We added the second DAR module. Also, the MI (3.346) has improved, but other indicators have declined. This proves that the DAR module is conducive to MI and it is adverse to other indicators. We add the TFM module between two DAR modules to balance local and global feature extraction. The results indicate that the SD and Q_{abf} improve to 79.894 and 0.508, respectively. When the second TF block is equipped, all the indicators are improved steadily. When the DCCA is equipped, the MI, SD, and VIF are improved. Overall, a network equipped with the DCCA and TF modules can extract local feature information, multiscale information, and global contextual information from source images.

4.7 Failure Cases

Fig. 4 shows experiment results with poor fusion effect. In the first and second columns, images (a) and (b) are the original CT and MRI images, respectively. The third column shows the fusion results for each row of original images. From an image fusion standpoint, images

Table 4 Ablation studies on vital components. (The best results are marked in bold)

Methods	MI ↑	SD ↑	VIF ↑	Q_{abf} ↑
baseline	3.232	80.311	0.507	0.535
baseline + DAR	3.346	66.333	0.509	0.510
baseline + DAR + DAR	3.353	64.305	0.498	0.495
baseline + DAR +TF + DAR	3.302	79.894	0.451	0.508
baseline + DAR + TF + TF +DAR	3.457	83.733	0.496	**0.536**
baseline + DAR + TF + TF + DAR + DCCA	**3.468**	**83.909**	**0.520**	0.508



(a) CT (b) MRI (c) Fusion

Figure 4 Illumination of failure cases.

(c) successfully integrate information from the original images. However, in the context of clinical medicine, critical MRI details are obscured by the CT images during the fusion process. The reason is that the fusion rule selects the fusion site from the pixel values which will lead to fusing too much CT luminance information. Therefore, in the subsequent study, we will modify the fusion rules to place certain restrictions on the brightness information of CT images.

5 Conclusion

In this work, we propose the DCATNet for medical image fusion. The DCATNet is constructed using DCCA and TF. Specifically, the proposed module DCCA extracts local feature information through CPA and leverages dilated convolutions to extract multi-scale information from the source images. The TF module obtains global contextual information, which achieves various information complementarities to form a fused image with comprehensive information. This experiment utilized two mainstream databases, namely CT-MRI and PET-MRI. Both datasets are used to measure the fusion results. Numerous experimental results show that DCATNet performs well in both qualitative and

quantitative evaluations and prove the superiority of our method. Ablation experiments have also confirmed the effectiveness of DCATNet. However, the DCATNet has no advantage in computational complexity, and the MRI details are obscured by the CT images in some fusion images. In future work, more efforts are expected on the lightweight network design and the CT and MRI image fusion strategy.

References

[1] Al-Fahdawi, S., Al-Waisy, A. S., Zeebaree, D. Q., Qahwaji, R., Natiq, H., Mohammed, M. A., Nedoma, J., Martinek, R. & Deveci, M. (2024), 'Fundus-deepnet: Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images', *Information Fusion* **102**, 102059.

[2] Cheng, C., Xu, T. & Wu, X.-J. (2023), 'Mufusion: A general unsupervised image fusion network based on memory unit', *Information Fusion* **92**, 80–92.

[3] Daneshvar, S. & Ghassemian, H. (2010), 'Mri and pet image fusion by combining ihs and retina-inspired models', *Information fusion* **11**(2), 114–123.

[4] Dinh, P.-H. (2023), 'Medical image fusion based on enhanced three-layer image decomposition and chameleon swarm algorithm', *Biomedical Signal Processing and Control* **84**, 104740.

[5] Faragallah, O. S., Muhammed, A. N., Taha, T. S. & Geweid, G. G. (2021), 'Pca based svd fusion for mri and ct medical images', *Journal of Intelligent & Fuzzy Systems* **41**(2), 4021–4033.

[6] Gao, Y., Ma, S., Liu, J., Liu, Y. & Zhang, X. (2021), 'Fusion of medical images based on salient features extraction by pso optimized fuzzy logic in nsst domain', *Biomedical Signal Processing and Control* **69**, 102852.

[7] Guo, Z., Li, X., Huang, H., Guo, N. & Li, Q. (2018), Medical image segmentation based on multi-modal convolutional neural network: Study on image fusion schemes, *in* '2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)', IEEE, pp. 903–907.

[8] Hou, R., Zhou, D., Nie, R., Liu, D. & Ruan, X. (2019), 'Brain ct and mri medical image fusion using convolutional neural networks and a dual-channel spiking cortical model', *Medical & biological engineering & computing* **57**, 887–900.

[9] Kadhim, D. A. & Mohammed, M. A. (2024), 'A comprehensive review of artificial intelligence approaches in kidney cancer medical images diagnosis, datasets, challenges and issues and future directions', *International Journal of Mathematics, Statistics, and Computer Science* **2**, 199–243.

[10] Li, H., Wu, X.-J. & Kittler, J. (2021), 'Rfn-nest: An end-to-end residual fusion network for infrared and visible images', *Information Fusion* **73**, 72–86.

[11] Liu, Y., Chen, X., Peng, H. & Wang, Z. (2017), 'Multi-focus image fusion with a deep convolutional neural network', *Information Fusion* **36**, 191–207.

[12] Ma, J., Tang, L., Xu, M., Zhang, H. & Xiao, G. (2021), 'Stdfusionnet: An infrared and visible image fusion network based on salient target detection', *IEEE Transactions on Instrumentation and Measurement* **70**, 1–13.

[13] Ma, J., Yu, W., Liang, P., Li, C. & Jiang, J. (2019), 'Fusiongan: A generative adversarial network for infrared and visible image fusion', *Information fusion* **48**, 11–26.

[14] Ma, J., Zhang, H., Shao, Z., Liang, P. & Xu, H. (2020), 'Ganmcc: A generative adversarial network with multiclassification constraints for infrared and visible image fusion', *IEEE Transactions on Instrumentation and Measurement* **70**, 1–14.

[15] Nair, R. R., Singh, T., Sankar, R. & Gunndu, K. (2021), 'Multi-modal medical image fusion using lmf-gan-a maximum parameter infusion technique', *Journal of Intelligent & Fuzzy Systems* **41**(5), 5375–5386.

[16] Ram Prabhakar, K., Sai Srikar, V. & Venkatesh Babu, R. (2017), Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 4714–4722.

[17] Song, W., Zeng, X., Li, Q., Gao, M., Zhou, H. & Shi, J. (2024), 'Ct and mri image fusion via multimodal feature interaction network', *Network Modeling Analysis in Health Informatics and Bioinformatics* **13**(1), 13.

[18] Tang, W., He, F., Liu, Y., Duan, Y. & Si, T. (2023), 'Datfuse: Infrared and visible image fusion via dual attention transformer', *IEEE Transactions on Circuits and Systems for Video Technology*.

[19] Wang, C., Yang, G., Papanastasiou, G., Tsaftaris, S. A., Newby, D. E., Gray, C., Macnaught, G. & MacGillivray, T. J. (2021*a*), 'Dicyc: Gan-based deformation invariant cross-domain information fusion for medical image synthesis', *Information Fusion* **67**, 147–160.

[20] Wang, J., Li, W., Wang, Y., Tao, R. & Du, Q. (2023), 'Representation-enhanced status replay network for multisource remote-sensing image classification', *IEEE Transactions on Neural Networks and Learning Systems*.

[21] Wang, Z., Li, X., Duan, H., Su, Y., Zhang, X. & Guan, X. (2021*b*), 'Medical image fusion based on convolutional neural networks and non-subsampled contourlet transform', *Expert Systems with Applications* **171**, 114574.

[22] Xia, K.-j., Yin, H.-s. & Wang, J.-q. (2019), 'A novel improved deep convolutional neural network model for medical image fusion', *Cluster Computing* **22**, 1515–1527.

[23] Xu, H. & Ma, J. (2021), 'Emfusion: An unsupervised enhanced medical image fusion network', *Information Fusion* **76**, 177–186.

[24] Xu, H., Gong, M., Tian, X., Huang, J. & Ma, J. (2022), 'Cufd: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition', *Computer Vision and Image Understanding* **218**, 103407.

[25] Xu, H., Ma, J., Jiang, J., Guo, X. & Ling, H. (2020), 'U2fusion: A unified unsupervised image fusion network', *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[26] Xu, H., Zhang, H. & Ma, J. (2021), 'Classification saliency-based rule for visible and infrared image fusion', *IEEE Transactions on Computational Imaging* **7**, 824–836.

[27] Xydeas, C. S., Petrovic, V. et al. (2000), 'Objective image fusion performance measure', *Electronics letters* **36**(4), 308–309.

[28] Yang, Y., Tong, S., Huang, S., Lin, P. et al. (2014), 'Log-gabor energy based multimodal medical image fusion in nsct domain', *Computational and mathematical methods in medicine*.

[29] Zebari, N. A., Mohammed, C. N., Zebari, D. A., Mohammed, M. A., Zeebaree, D. Q., Marhoon, H. A., Abdulkareem, K. H., Kadry, S., Viriyasitavat, W., Nedoma, J. et al. (2024), 'A deep learning fusion model for accurate classification of brain tumours in magnetic resonance images', *CAAI Transactions on Intelligence Technology*.

[30] Zhai, W., Song, W., Chen, J., Zhang, G., Li, Q. & Gao, M. (2023), 'Ct and mri image fusion via dual-branch gan', *International Journal of Biomedical Engineering and Technology* **42**(1), 52–63.

[31] Zhang, H. & Ma, J. (2021), 'Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion', *International Journal of Computer Vision* pp. 1–25.

[32] Zhang, H., Le, Z., Shao, Z., Xu, H. & Ma, J. (2021*a*), 'Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion', *Information Fusion* **66**, 40–53.

[33] Zhang, H., Xu, H., Tian, X., Jiang, J. & Ma, J. (2021*b*), 'Image fusion meets deep learning: A survey and perspective', *Information Fusion* **76**, 323–336.

[34] Zhang, H., Yuan, J., Tian, X. & Ma, J. (2021*c*), 'Gan-fm: Infrared and visible image fusion using gan with full-scale skip connection and dual markovian discriminators', *IEEE Transactions on Computational Imaging* **7**, 1134–1147.

[35] Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R. & Van Gool, L. (2023), Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 5906–5916.

[36] Zhou, T., Ruan, S. & Canu, S. (2019), 'A review: Deep learning for medical image segmentation using multi-modality fusion', *Array* **3**, 100004.