J. Meas. Sci. Instrum., 2025, 16(1): 26-36. http://jmsi.publish.founderss.cn
DOI: 10.62756/jmsi.1674-8042.2025003 jmsi@nuc.edu.cn

Multimodal medical image fusion based on mask optimization and parallel attention mechanism

DI Jing*, LIANG Chan, GUO Wenqing, LIAN Jing

School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

***Corresponding author:** DI Jing (dijing_lzjtu@163.com)

Received: May 16, 2024 **Revised:** August 3, 2024 **Accepted:** September 1, 2024

Abstract: Medical image fusion technology is crucial for improving the detection accuracy and treatment efficiency of diseases, but existing fusion methods have problems such as blurred texture details, low contrast, and inability to fully extract fused image information. Therefore, a multimodal medical image fusion method based on mask optimization and parallel attention mechanism was proposed to address the aforementioned issues. Firstly, it converted the entire image into a binary mask, and constructed a contour feature map to maximize the contour feature information of the image and a triple path network for image texture detail feature extraction and optimization. Secondly, a contrast enhancement module and a detail preservation module were proposed to enhance the overall brightness and texture details of the image. Afterwards, a parallel attention mechanism was constructed using channel features and spatial feature changes to fuse images and enhance the salient information of the fused images. Finally, a decoupling network composed of residual networks was set up to optimize the information between the fused image and the source image so as to reduce information loss in the fused image. Compared with nine high-level methods proposed in recent years, the seven objective evaluation indicators of our method have improved by 6%—31%, indicating that this method can obtain fusion results with clearer texture details, higher contrast, and smaller pixel differences between the fused image and the source image. It is superior to other comparison algorithms in both subjective and objective indicators.

Key words: multimodal medical image fusion; binary mask; contrast enhancement module; parallel attention mechanism; decoupling network

0 Introduction

Multimodal medical imaging leverages diverse diagnostic technologies to enhance clinical assessments. Magnetic resonance imaging (MRI) provides high-resolution visuals and detailed soft tissue characterization. In contrast, positron emission tomography (PET) and single-photon emission computed tomography (SPECT) images capture cellular and tissue activity. The integration of these modalities offers a comprehensive view of various pathological states, facilitating precise diagnoses and treatment strategies. This fusion not only mitigates the risks of misdiagnosis and surgical errors, but also elevates the overall safety and efficacy of medical interventions.

Image fusion techniques have significantly advanced across various domains, leading to the development of numerous algorithms. Among these, multiscale transform (MST) methods such as wavelet transform[1], pyramid transform[2], non-subsampled shearlet transform (NSST)[3], and non-subsampled contourlet transform (NSCT)[4] are prominent. These methods decompose images into high and low frequency bands, effectively capturing contour details. However, they rely on multi-step decomposition, resulting in high complexity. In recent years, deep learning has revolutionized image fusion, often outperforming traditional techniques[5]. Key categories include convolutional neural networks (CNNs)[6,7], autoencoders (AEs)[8], and generative adversarial networks (GANs)[9]. CNNs excel in feature extraction, fusion, and reconstruction, optimizing fusion results through network design and loss functions. AEs utilize similar principles for feature handling, focusing on efficient fusion strategies. GANs employ generators and discriminators to refine fusion quality, ensuring high fidelity to the source images. Several notable studies have expanded on these foundations. Xu et al.[10] proposed a unified unsupervised image fusion network to automatically estimate the importance of the corresponding source image through feature extraction and information measurement. Although more information in the source image is retained,

© The Author(s) 2025. Published by Publishing Center of North University of China. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

the contrast problem of medical images is neglected in this process, resulting in poor overall brightness of the fusion result. Zhang et al.[11] proposed a fusion decomposition model for CNNs, which uses convolutional blocks in both fusion and decomposition phases, reduceingg the algorithmic complexity but extracting less feature information, resulting in more soft tissue and contour information. Liu et al.[12] proposed a coupled contrast learning network that retains the main features of the two modalities to remove the redundant information appearing in the fused image, and used a mask to highlight the contour information but ignored the overall brightness information of the image, resulting in a lower overall clarity of the fused image. Liu et al.[13] proposed a multimodal image fusion with spatial attention mechanism and wavelet constraints by combining the spatial attention module (SAM) into the generator to obtain a spatial attention map using the attention map to force the discrimination of the multimodal image to focus on the target region, and introducing wavelet constraints into the generator to force the generator to incorporate more detailed information. Although this method improves the texture detail features through the spatial attention mechanism, it does not pay attention to the luminance information of the image in this process, which makes the edge contour and soft tissue information in the fusion result dark, and the visual effect of the human eye is poor.

To address blurred textures and low contrast in multimodal medical image fusion, this study introduces a novel method integrating mask optimization and a parallel attention mechanism. The technique uses binary mask transformation for accurate texture detail capture and a triple-path input system for improved detail extraction. Image enhancement and detail preservation modules enhance brightness and clarity, while the parallel attention mechanism dynamically adjusts fusion weights to highlight key features. A residual-based decoupling network minimizes information loss, ensuring the fused output retains rich texture details.

1 Methods

The multimodal medical image fusion model presented in this paper features a dual-component architecture: the fusion network and the decoupling network, as illustrated in Fig. 1. The fusion network employs two unimodal medical images, a binary mask and source images, as inputs. These are processed through a multilayer CNN to extract image features, enhanced by a contrast enhancement module and a detail preservation module. Fusion of the final images is achieved via a compensatory attention mechanism. The decoupling network utilizes a full convolutional network to ensure that the fused images retain enhanced texture and achieve pixel-level detail precision.

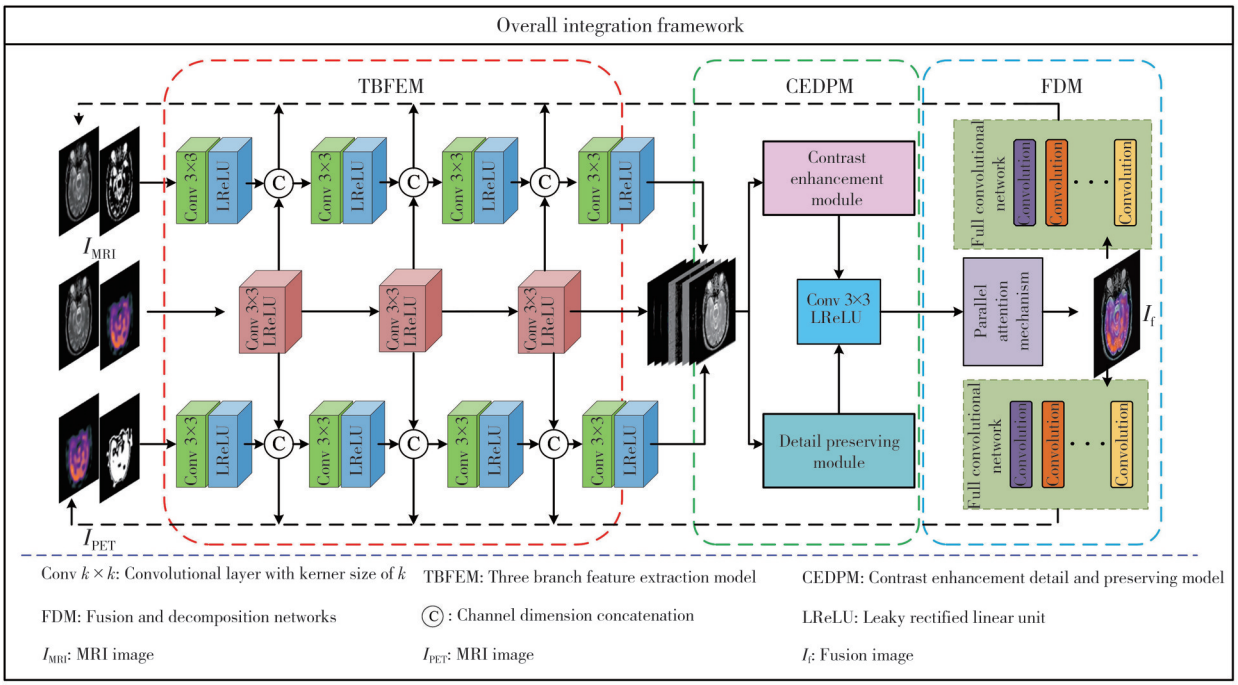

Fig. 1 Mask optimization and parallel attention mechanism network process

1.1 Binary mask

Multimodal medical images, combining functional and anatomical data, require precise texture and contour delineation for effective diagnosis. Anatomical images provide rich texture and structural details, but previous

deep learning frameworks often failed to adequately preserve these elements during fusion, leading to diminished clarity and hampering accurate lesion detection. Traditional approaches that directly input MRI and PET/SPECT modalities into a primary network for feature extraction tend to exacerbate detail loss. Therefore, to reduce the information loss of source images during fusion, a mask is introduced to enhance the texture features of fused images. In this work, the binary mask is constructed by Otsu algorithm, and the specific steps are as follows. For image $t(x,y)$, firstly, its adjustment threshold is set to T, so as to distinguish the texture target from the background that want to be extracted from the medical image. Secondly, the ratio of the number of target pixels to that of the whole image is set to ω_0, with the average grey scale of σ_0, the ratio of the number of background pixels to the whole image is set to ω_1, with the average grey scale of σ_1, the average grey scale of the whole image is set to σ, and the inter-class variance is g. Finally, assuming that the size of the input image is $M\times N$, the number of pixels in the image whose grey value is less than the final threshold T is N_0, and the number of pixels whose pixel grey value is greater than the final threshold T is N_1, the above computational formulas can be expressed as

$$\omega_0 = \frac{N_0}{M \times N}, \tag{1}$$

$$\omega_1 = \frac{N_1}{M \times N}, \tag{2}$$

$$N_0 + N_1 = M \times N, \tag{3}$$

$$\omega_0 + \omega_1 = 1, \tag{4}$$

$$\sigma = \omega_0\sigma_0 + \omega_1\sigma_1, \tag{5}$$

$$g = \omega_0\omega_1(\sigma_0 - \sigma_1)^2. \tag{6}$$

Following the outlined steps, distinct binary masks for three groups of medical images are demonstrated in Fig.2. These masks effectively separate the texture targets and background information, preserving the original texture details and contour structures of the source images.

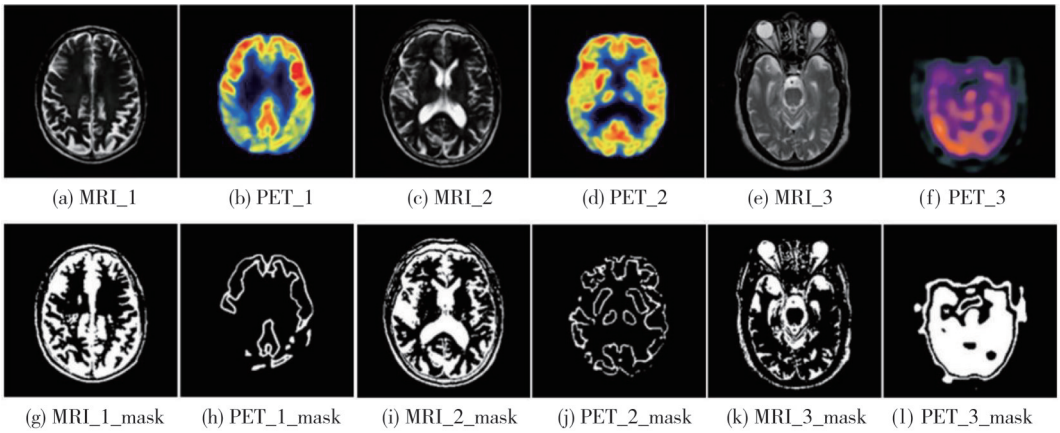

Fig. 2 Three series of binary mask images

1.2 Feature extraction

Upon processing the source image with the Otsu algorithm, the input features are segmented into three distinct paths. Two paths integrate MRI and PET/SPECT source images alongside their corresponding binary masks, which compensates for the loss of texture details and contour information typically incurred during feature extraction. The inclusion of binary masks enriches structural data, significantly enhancing the structural clarity of the fused image to align more closely with human visual perception. In feature extraction, texture details are derived using four convolutional blocks, each with a kernel size of 3×3 and Leaky ReLU (LReLU) activation function. While the binary mask mitigates information loss, the initial paths through the CNN still risk losing structural information from MRI and functional data from PET/SPECT images. To address this, a third path is introduced, merging both source images to restore and optimize structural and functional integrity. This pathway employs three additional convolutional blocks with a kernel size of 3×3 and LReLU activation, strategically enhancing feature information and further augmenting the texture detail extraction in the final image.

1.3 Feature fusion

After extracting features, the feature maps from the three pathways are cascaded across channels to integrate diverse image details effectively. Recognizing that diagnostic image evaluation requires not only high clarity in texture and contours but also suitable overall brightness, a contrast enhancement module is adopted for incorporated post-feature extraction to improve image luminance.

Additionally, medical imaging often necessitates emphasizing significant information while minimizing redundant data. To achieve this, a parallel attention mechanism is introduced. This mechanism dynamically adjusts weights to prioritize critical image details and employs both channel and spatial attention strategies to retain these significant features. This process ensures optimal utilization of valuable information, culminating in the production of a well-fused image that meets diagnostic needs.

1.3.1 Contrast enhancement module

The contrast enhancement module (CEM) is shown in Fig. 3, which mainly consists of a multiscale layer, a residual block, and a weight block.

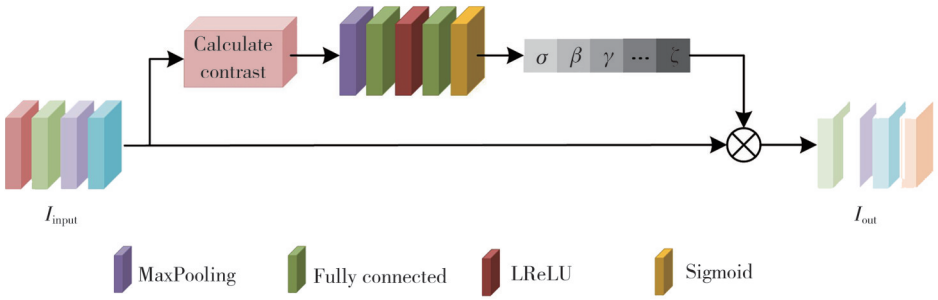

Fig. 3 Contrast enhancement module

The multiscale layer uses four convolutional layers with kernel sizes of 1×1, 3×3, 5×5, and 7×7 to capture the multiscale depth features after the feature cascade to improve the ability to capture global dependencies, and the activation function of the convolutional layer is LReLU. In the contrast block, the standard deviation of each position in the residual flow features is calculated as

$$\sigma_{ij}=\frac{1}{2r+1}\sqrt{\sum_{-r\leqslant p,q\leqslant r}(\varphi_m(i+p,j+q)-\mu_{ij})^2}, \quad (7)$$

where r is the window size, and $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean and standard deviation of the centre at (i,j), respectively. Thus, the contrast map is output after calculating the standard deviation, followed by the maximum pooling layer, the fully connected layer with LReLU and the fully connected layer with sigmoid. The final output of the residual network is an activation vector representing the feature contrast weights. Subsequently, the original features and the final output channel of the residual network are multiplied to generate enhanced feature information.

CEM enhances the brightness and contrast of images, making details clearer and more visible. First, it uses convolutional layers with different kernel sizes (1×1, 3×3, 5×5, 7×7) to capture multiscale depth features, improving the ability to capture global dependencies. Next, it calculates the standard deviation of residual flow features to highlight high-contrast areas, thereby better showcasing details. Finally, it generates an activation vector representing feature contrast weights by computing the contrast map, which is then multiplied with the original features to enhance the final feature information, significantly improving the visibility of details and overall processing effectiveness.

1.3.2 Detail preserving module

To enhance the preservation of structural details from the source image, this study introduces the detail preserving module (DPM), as depicted in Fig.4.

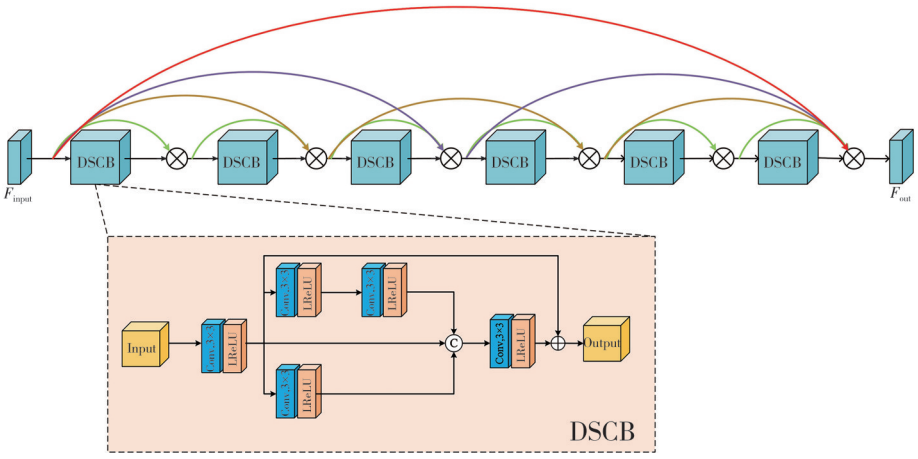

Fig. 4 Detail preserving module

The DPM comprises six depthwise separable convolution blocks (DSCBs) interconnected with jump connections at varying intervals. Specifically, these jump connections are strategically placed at short, medium, and long distances across one, two, three, and six DSCBs, respectively. This design allows the DPM to capture features at multiple scales, facilitating comprehensive feature extraction from different levels and integrating a residual network structure to boost performance.

The module initiates with a 3×3 convolutional layer followed by an LReLU activation layer. Subsequently, three parallel paths extract features across diverse receptive fields. The outputs of these paths are concatenated in the channel dimension, followed by a convolutional layer with a kernel size off 3×3 to regulate the channel output count to match the input. A residual structure is established by summing the output with the input, optimizing learning and improving the model's capability to retain intricate details.

Overall, the use of dense connection structures in the DPM aims to retain more useful information, not just detail information. This information includes edges, textures, and brightness features. By enhancing and fusing these features, higher-quality images that better meet practical application needs can be generated.

1.3.3 Parallel attention mechanisms

This study introduces a parallel attention mechanism module, inspired by both channel and spatial attention mechanisms, to enhance path information delivery and global feature modeling in image fusion. As depicted in Fig. 5, the module incorporates refined versions of both attention mechanisms.

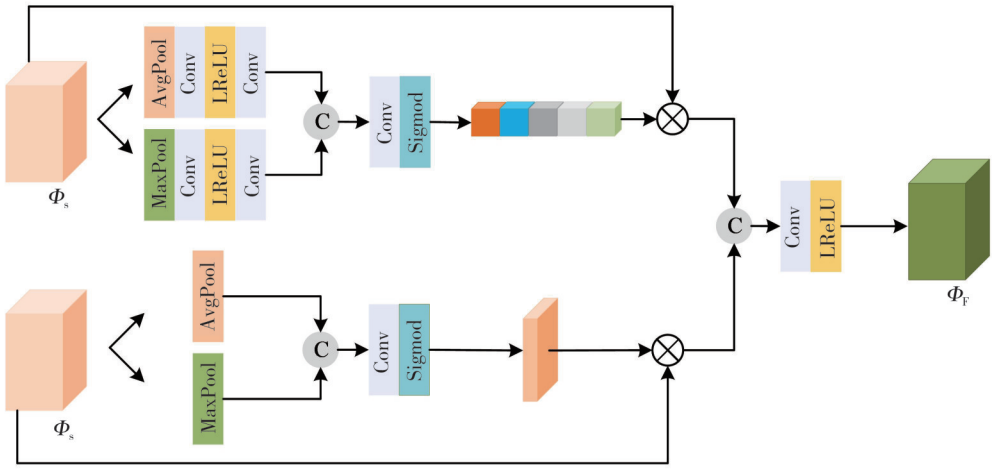

Fig. 5 Cascading attention mechanism model

The channel attention mechanism at the upper level initiates by merging features through both average and maximum pooling operations. Maximum pooling is utilized to extract predominant information from the most significant parts of the feature map, aiding in better target size and shape recognition. Additionally, average pooling, provides a comprehensive capture of the overall feature map information and aids in image smoothing. The pooled outputs are then concatenated and processed through two 3×3 convolutional layers followed by an LReLU activation layer, and subsequently passed through additional convolutional and sigmoid activation layers. The lower-level enhanced spatial attention mechanism takes the processed feature information as input, using both averaging and maximum pooling to generate spatial feature maps. These maps are then concatenated and directed through convolutional and sigmoid activation layers to derive initial spatial weighting coefficients. Finally, the outputs from the upper and lower attention mechanisms are combined using a 3×3 convolutional layer and an LReLU activation layer, resulting in an output that maximizes feature utilization for image fusion. This dual-layered attention framework allows for separate and effective learning of feature maps using improved channel and spatial attention techniques.

1.4 Decoupling network

The decoupling network in this study is designed to minimize the loss of texture and contour information in the fused image by reprocessing it and comparing it against the source image data. As detailed in this paper, the fusion results are input directly into a fully convolutional network that maintains the original image size through each layer, which enhances outcomes in deeper network levels. To maximize the extraction of feature information from the input image, each segment employs dual convolutions with kernel sizes of 1×1 and 3×3, coupled with LReLU activation function, as

illustrated in Fig.6. Notably, the output from the fourth layer of the feature extraction phase is fed back into the first layer of the decomposition network. This technique addresses the issue of information loss during the decomposition process by reintegrating critical data from the deeper layers, thereby enriching the information about the original image and narrowing the disparity with the source image.

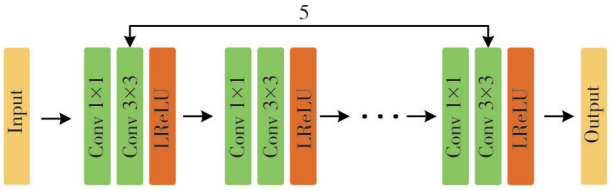

Fig. 6 Single branch decomposition network model

1.5 Loss function

The network architecture in this study has two components: an image fusion component and an image decoupling component. The fusion network is fused into a single image by feature extraction, while the decoupling network aims to separate and refine the features of the fused image to improve the image quality and ensure the retention of key information. The overall loss function consists of the fusion loss L_{fusion} and decoupling loss L_{decouple}, and can be expressed as

$$L_{\text{total}} = L_{\text{fusion}} + L_{\text{decouple}}. \tag{8}$$

1.5.1 Fusion loss

This study is to transform the medical image fusion problem by extracting and reconstructing the gradient and intensity information, so the fusion loss also consists of these two parts. The loss function can be expressed as

$$L_{\text{fusion}} = \beta L_{\text{gradient}} + L_{\text{intensity}}, \tag{9}$$

where β is used to balance the intensity and gradient terms.

Intensity loss quantifies the difference between the fused image and the source image in terms of pixel intensity, and the mean squared error (MSE) loss function is simple, fast and effective in calculation, which is calculated by

$$L_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}\left\| Y_i - \hat{Y}_i \right\|^2, \tag{10}$$

where n is the number of samples, Y_i is the true value of sample i, and $\hat{Y}_i$ is the predicted value of the model for sample i.

Intensity loss affects the contrast of the fused image and MSE reduces the error between the generated image and the original image. This effectively adjusts the contrast information between the fused image and the two input images. Therefore, the intensity degree loss can be expressed as

$$L_{\text{intensity}} = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left\| I_{\text{fused}_{i,j}} - I_{1_{i,j}} \right\|_2^2 + \alpha\left\| I_{\text{fused}_{i,j}} - I_{2_{i,j}} \right\|_2^2, \tag{11}$$

where i and j denote the pixels in the ith row and jth column weight map, respectively; H and W represent the height and width of the image, respectively; I_1 and I_2 are the source images; I_{fused} is the fused image; and $\|\cdot\|^2$ denotes the square of the Euclidean distance, which is used to quantify difference between the two images.

Gradient loss is concerned with the difference between the fused image and the source image in terms of edge and texture information, and it is calculated by

$$L_{\text{gradient}} = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1} S_{1i,j}\left\| \nabla I_{\text{fused}_{i,j}} - \nabla I_{1_{i,j}} \right\|_2^2 + S_{2i,j}\left\| \nabla I_{\text{fused}_{i,j}} - \nabla I_{2_{i,j}} \right\|_2^2, \tag{12}$$

where i and j denote the pixels in the ith row and jth column weight map, respectively; H and W represent the height and width of the image, respectively; I_1 and I_2 are the source images; I_{fused} is the fused image; $\|\cdot\|^2$ denotes the square of the Euclidean distance; ∇ is the gradient operator; and $S_{i,j}$ is usually computed by using Sobel's operator, and it is the adaptive weight, reflecting the importance of the gradient information.

1.5.2 Decoupling loss

For decoupling networks, the requirement that the decoupling result of the fused image should be as consistent as possible with the source image can be expressed as the difference between the fused image and the decoupled image. A common form is the MSE, which is used to quantify the difference between the decoupled image and the fused image.

This is obtained by substituting the fused image and the two decoupled images into

$$L_{\text{decouple}} = \frac{1}{HW}\sum_{i=0}^{H-1}\sum_{j=0}^{W-1}\left\| (I_{1\_\text{de}_{i,j}} - I_{1_{i,j}}) \right\|_2^2 + \left\| (I_{2\_\text{de}_{i,j}} - I_{2_{i,j}}) \right\|_2^2, \tag{13}$$

where I_{1_de} and I_{2_de} are the decoupling results of the fused images, respectively; I_1 and I_2 are the source images; $\|\cdot\|^2$ denotes the square of the Euclidean distance; i and j denote pixels in the ith row and jth column weight map; and H and W represent the height and width of the image, respectively.

2 Experimental results and discussion

The experimental setup in this study utilized a Windows 11 64-bit) operating system on an Intel (R) Core (TM) i7-

12700H 2.30 GHz laptop with 16 GB of RAM. This hardware configuration supports the performance assessment and validation of the proposed image fusion model. Utilizing the Harvard University Medical Library dataset (http://www.med.harvard.edu/AANLIB/home.html), 180 pairs of 256×256 pixels medical images were segmented into 35 280 training blocks measuring 120×120 pixels each. For training, in order to obtain more training data, we adopted the expansion strategy of tailoring and decomposition. The training regimen included 30 iterations with a batch size of 32 and an initial learning rate of 1×10^{-4}. The Adam optimization algorithm was selected for its adaptiveness. The loss function integrated gradient, intensity, and decoupling losses with respective proportions of 10, 1, and 1, as determined through extensive experiment. Since the proposed method is a fully convolutional network, the source images do not need to be cropped into small patches with the same size (in pixels) as the training data during the test phase. In other words, the test is performed on the original size of source images.

2.1 Objective evaluation indicators

To validate the effectiveness of the algorithms presented in this paper, several objective metrics are employed to evaluate the quality of the fused images. These include average gradient (AG)[14], information entropy (EN)[15], spatial frequency (SF)[16], variance (SD)[14], correlation coefficient (CC)[17], mutual information (MI)[18], and sum of correlation of differences (SCD)[19]. AG and EN assess image clarity and informational content, respectively, with higher values indicating richer and more comprehensive information extraction. SD measures the distribution and contrast within the fused image, providing statistical representation of its quality. MI quantifies the similarity between the fused image and source images, while CC, nearing a value of 1, indicates high similarity in feature information between the images. SF evaluates the greyscale fidelity of the fused image compared to the source, with higher values indicating closer alignment. Lastly, SCD assesses the correlation level between the information conveyed to the fused image and its corresponding source images, reflecting the effectiveness of the fusion process.

2.2 Comparison of experimental results

In order to verify the validity and practicality of the proposed methodology, subjective visual evaluation and objective data comparisons were performed with nine comparable methods: LEGFF[20], CNP[21], SDNet[11], DTNP[22], MATR[23], MDL[24], CNN[25], MLEPF[26] and SwinFusion[27].

2.2.1 Subjective visual evaluation

Fig.7 illustrates the MRI-SPECT image of a "Grade II astrocytoma."

The MRI and SPECT/PET source images are shown in Fig.7(a) and Fig.7(b), respectively. The fusion results from SDNet, MATR, and SwinFusion methods exhibit blurriness with missing texture details. MLEPF results display a purple block effect in the edge contours, introducing non-original color information. LEGFF and CNN methods fail to retain MRI soft tissue details and display poor bone edge contours, while MDL method shows weak soft tissue texture and contrast. Conversely, CNP and DTNP methods capture more detailed information but suffer from blurring at target edges. In contrast, our proposed method (Fig. 7(l)) effectively retains SPECT/PET color information and enhances soft tissue brightness and texture detail from MRI images, producing clearer results.

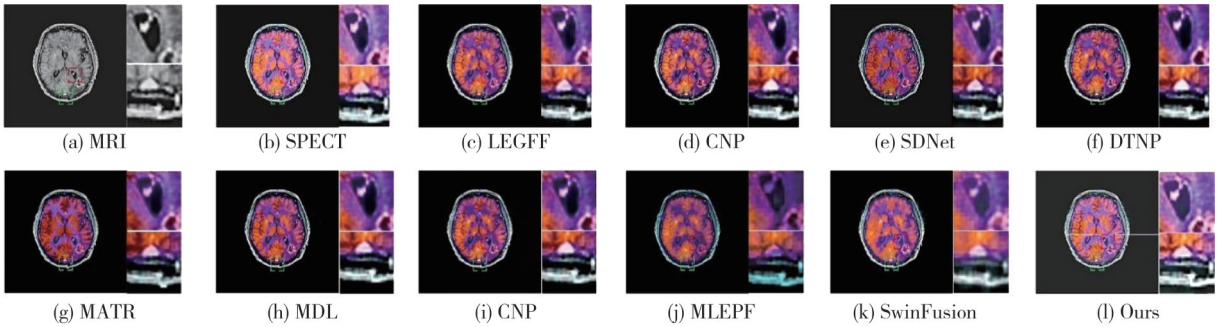

Fig. 7 Comparison of fusion and details of MRI SPECT images of grade II astrocytoma

Fig.8 presents an MRI-SPECT image of "acute stroke disease". LEGFF method results in blurred images with significant edge loss. MATR and CNP methods also lose critical detail, hindering lesion area observation. MLEPF produces artifacts in soft tissue areas, affecting clarity and image quality. Although SDNet and CNN methods preserve clear edge information, their overall images are dark. MDL, DTNP, and SwinFusion methods generally offer better detail but suffer from low contrast and slight edge blurring. In comparison, our method maintains

complete edge details and contours with clear, detailed contrast.

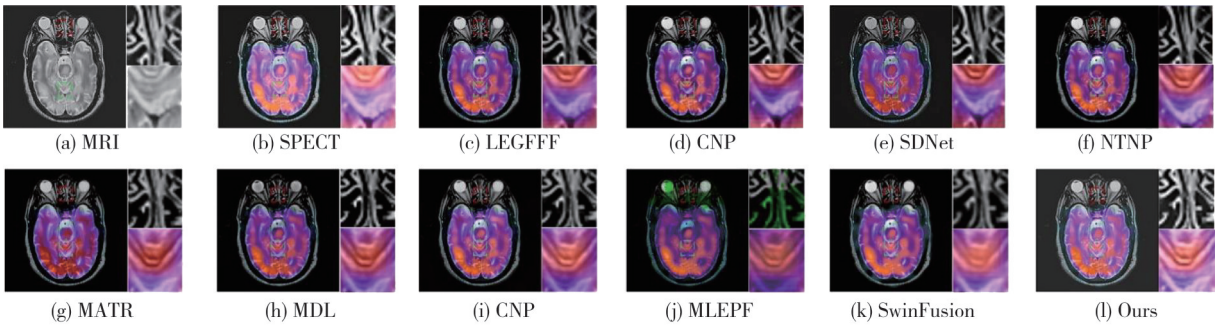

Fig. 8 Comparison of MRI SPECT image fusion and details in mild acute stroke disease

Fig.9 depicts MRI-PET images of "mild Alzheimer's disease". LEGFF method shows artifacts and blurred contours. SDNet, MATR, and MDL methods result in darker, blurred images. MLEPF method exhibits a purple hue in soft tissues. CNP and CNN methods are darker in the middle, capturing PET color information but failing to integrate MRI soft tissue details adequately. DTNP and SwinFusion methods significantly lose contour information on the right edge and fail to display original MRI contours. In companison, our fusion algorithm generates the images with clear soft tissue textures and contours, closely aligning with human visual characteristics.

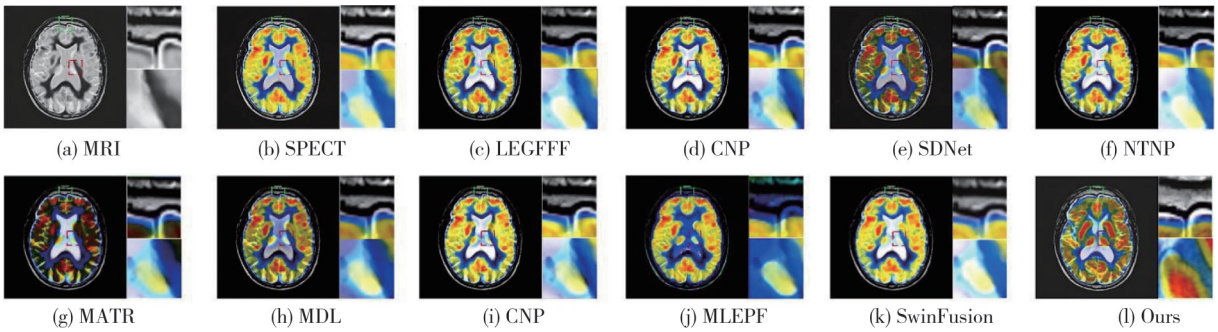

Fig. 9 Comparison of MRI PET image fusion and details in mild Alzheimer's disease

2.2.2 Objective evaluation

The fusion results of the three groups of medical images show that the fused images obtained by the multimodal medical image fusion strategy proposed in this study are highly consistent with the visual characteristics of the human eye, and the details are more fully extracted and the colours are more natural. As can be seen from Tables 1—3, the performance of AG, EN, SF, MI and SCD are all better, but the performance of CC is slightly weaker than that of LEGFF in "normal brain" and SD in "mild Alzheimer's disease", which is slightly weaker than that of SwinFusion.

Table 1 Objective evaluation indicators for MRI SPECT image fusion of grade II astrocytoma images

Method	AG	EN	SF	MI	CC	SD	SCD
LEGFF	9.235 8	2.963 8	50.603 5	2.530 5	0.713 1	7.963 5	1.523 1
CNP	9.581 1	2.693 5	53.122 1	2.478 5	0.701 7	7.957 9	1.394 7
SDNet	9.740 2	3.362 9	51.786 0	2.502 9	0.598 7	7.232 9	0.199 5
DTNP	9.628 4	2.726 4	53.192 3	2.476 2	0.696 5	7.953 6	1.413 1
MATR	8.145 7	2.752 6	44.818 2	2.529 4	0.599 9	7.535 2	0.511 8
MDL	8.584 3	2.722 7	49.240 4	2.434 5	0.688 1	7.731 4	1.279 1
CNN	8.885 5	3.269 8	48.690 3	2.510 4	0.686 8	7.713 5	1.156 1
MLEPF	7.185 2	3.027 1	41.120 5	2.412 1	0.725 1	7.325 7	0.873 7
SwinFusion	6.264 2	3.125 4	34.673 8	2.503 7	0.654 9	7.996 1	1.535 2
Ours	**9.995 8**	**3.595 3**	**53.365 8**	**2.576 8**	**0.735 6**	**8.160 3**	**1.652 8**

Table 2 Objective evaluation indicators for MRI SPECT image fusion in mild Acute Stroke disease

Method	AG	EN	SF	MI	CC	SD	SCD
LEGFF	9.351 8	5.320 7	39.249 0	3.529 9	0.524 1	0.616 9	8.058 7
CNP	9.509 6	4.284 3	41.959 9	3.556 2	0.724 5	0.586 3	8.095 7
SDNet	9.841 5	5.688 0	41.179 0	3.313 3	0.712 4	0.546 6	7.335 1
DTNP	9.456 0	4.313 0	41.790 1	3.549 5	0.729 1	0.591 7	8.103 7
MATR	8.375 6	4.488 0	36.606 9	**4.299 1**	0.744 4	0.503 4	7.906 9
MDL	8.863 5	4.516 8	39.888 0	3.083 6	0.715 7	0.640 3	7.716 7
CNN	8.786 2	5.265 2	37.590 7	3.696 5	0.767 8	0.632 5	8.045 8
MLEPF	6.695 1	4.343 8	32.619 3	3.175 4	0.507 9	0.731 9	7.166 2
SwinFusion	7.251 8	4.866 8	31.012 2	3.575 2	0.662 1	**0.653 0**	8.085 6
Ours	**11.903 4**	**5.797 8**	**51.536 6**	3.452 1	**0.774 5**	0.645 0	**8.617 6**

Table 3 Objective evaluation indicators for MRI PET image fusion in mild Alzheimer's disease

Method	AG	EN	SF	MI	CC	SD	SCD
LEGFF	9.125 4	4.327 3	46.987 5	3.464 9	**0.805 2**	8.817 3	1.183 4
CNP	9.895 4	3.586 5	50.387 5	3.481 0	0.746 5	9.014 0	1.549 1
SDNet	10.506 2	5.254 2	51.184 6	3.910 9	0.480 9	7.849 1	0.929 7
DTNP	10.021 8	3.650 3	50.663 6	3.543 6	0.734 2	9.503 3	1.561 1
MATR	8.615 7	3.406 6	45.914 5	3.919 6	0.273 6	8.061 2	0.833 3
MDL	9.275 6	3.874 0	47.136 4	3.647 7	0.713 8	8.452 9	1.441 0
CNN	8.944 2	3.540 1	45.695 8	3.187 7	0.715 7	9.236 3	1.569 3
MLEPF	7.698 9	4.249 4	40.098 9	3.897 4	0.716 9	9.465 7	1.150 0
SwinFusion	8.518 6	4.110 2	41.979 6	3.838 7	0.739 5	**9.624 5**	1.597 2
Ours	**11.156 7**	**5.376 2**	**54.635 4**	**3.929 8**	0.799 8	9.498 4	**1.599 6**

This is because LEGFF method pays too much attention to the overall contrast of the image and ignores the edge contour information of the image, while SwinFusion method mainly focuses on the information of PET image but ignores the texture details of MRI image, which results in the loss of the texture details of the image as a whole.

In contrast, the fusion process of this study focuses on the colour richness of PET images and SPECT images, as well as the edge contour information and soft tissue information of MRI images, and the fused images have clearer contour edges and more natural colours, which are more in line with the visual characteristics of the human eye, and can assist doctors in the rapid diagnosis and treatment of diseases.

2.3 Ablation experiment

To evaluate the effectiveness of the mask optimization, enhanced attention mechanism, contrast enhancement module, and decoupling network in our proposed image fusion method, four ablation experiments were conducted. These studies involved testing four modified versions of our network, each omitting one component: the binary mask (denoted as Mask), the improved attention mechanism (denoted as Attention), the contrast enhancement module (denoted as Contrast), and the decoupling network (denoted as Decouple). Additionally, a fifth experimental setup, integrating all components (denoted as Ours), was tested to observe comprehensive results. The configurations and results of these experiments are detailed in Table 4, as illustrated in Fig.10.

From Fig. 10, it can be seen that among the seven metrics putting all the network structures proposed in this study into the fusion framework the metrics obtain the optimal values, while the metrics are at a disadvantage compared to the full model after removing the individual structures, which proves the superiority of the mask optimization, the improved attentional mechanism, the contrast enhancement module, and the adversarial network used in proposed network.

Table 4 Different combinations of ablation experiments

Experiement	Mask	Attention	Contrast	Decouple
1	√	×	×	×
2	×	√	×	×
3	×	×	√	×
4	×	×	×	√
5	√	√	√	√

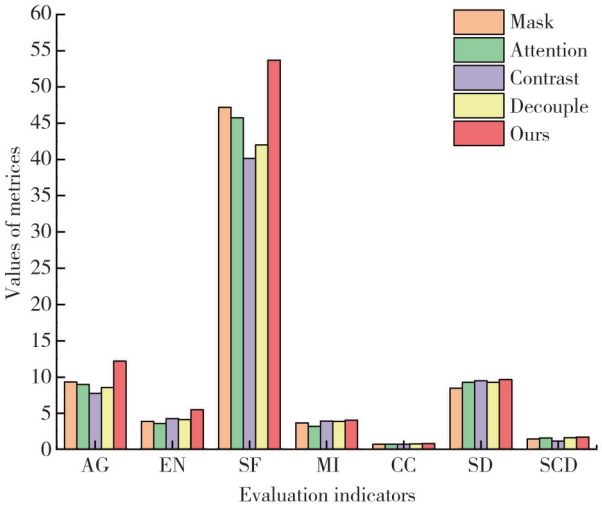

Fig. 10 Five different network structures in ablation experiment, 20 image mean bar charts

Fig.11 illustrates the four sets of fusion results generated by the different models, and it can be seen that the visual quality of the four images varies significantly, with the full model outperforming the other four sets of experiments in terms of detail information retention, texture contrast, and edge contours.

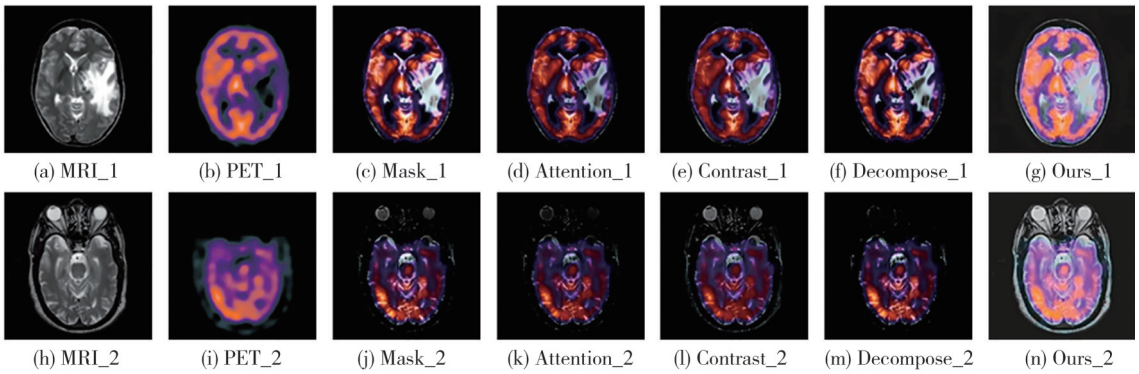

Fig. 11 Four sets of medical image fusion images by five different network structures for ablation experiment

2.4 Objective evaluation of ablation experiment

In order to evaluate the proposed method more comprehensively, 10 sets of MRI and PET images and 10 sets of MRI and SPECT images were randomly selected for testing and the average of 20 sets of data were taken, and the complexity and average running time were calculated to evaluate the method complexity and running efficiency.

Meanwhile, the proposed method was compared with other nine classical medical image fusion methods, and the specific comparison results are shown in Table 5.

Table 5 Average running time of different methods for 20 sets of images

Method	Time/s	Computational volume/GFLOPs
LEGFF	1.905 2	4.586 2
CNP	2.351 4	4.712 3
SDNet	**0.593 5**	2.458 7
DTNP	2.235 8	3.259 8
MATR	1.980 2	4.221 5
MDL	3.256 8	4.523 8
CNN	4.025 6	**2.014 2**
MLEDF	3.257 4	3.578 9
SwinFusion	1.742 1	2.541 4
Ours	1.702 5	2.531 4

In Table 5, GFLOPs means 10^9 FLOPs. It can be found that SDNet has the smallest computation and running time among the eight methods due to the fact that it has fewer network layers and only uses simple convolutional operations for feature extraction. However, this method is unable to extract feature information effectively, resulting in poor fusion results. Comparatively, the method in this study performs best in terms of average running time, and the number of parameters used in the testing phase is only 0.5×10^7, which can be considered as lightweight. Considering the overall metrics and method performance, the method in this study performs well in all aspects and can efficiently handle real-time fusion tasks.

3 Conclusions

In order to enhance the fusion effect of medical fusion images in human eye or machine vision features, a multimodal medical image fusion method with mask optimization and cascading attention mechanism is proposed by integrating the texture detail information in multimodal medical images. The method utilizes a binary mask to achieve effective extraction of texture detail features from multimodal medical images. The design of contrast enhancement module, detail preservation module, parallel attention mechanism and decoupling network achieves that the fused image contains more accurate pixel information and higher luminance information of the source image. Experiments show that the algorithm in this paper improves the objective evaluation metrics AG, EN, SF, MI, CC, SD, and SCD by an average of 18.78%, 29.46%, 19.08%, 8.65%, 17.91%, 6.55%, and 30.76%, respectively, compared with nine classical medical image fusion methods. This indicates that the method proposed in this paper achieves an advanced level of accuracy in both texture detail and contrast, offering significant value for doctors in treating diseases. Meanwhile, the extension of our method to the field of infrared and visible light image fusion is also applicable, which has significant practical application value in security and surveillance, medical diagnosis, as well as military and defense. However, the robustness issue of our image fusion framework still needs to be further improved. In future work, continue to explore how to effectively improve the fused image robustness and incorporate it into more advanced image fusion techniques.

Acknowledgement

This work was supported by Gansu Natural Science Foundation Programme (No. 24JRRA231), National Natural Science Foundation of China (No.62061023), and Gansu Provincial Education, Science and Technology Innovation and Industry (No.2021CYZC-04).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] CHAO Z, DUAN X G, JIA S F, et al. Medical image fusion *via* discrete stationary wavelet transform and an enhanced radial basis function neural network. Applied Soft Computing, 2022, 118: 108542.

[2] YU N N, LI J J, HUA Z. Decolorization algorithm based on contrast pyramid transform fusion. Multimedia Tools and Applications, 2022, 81(11): 15017-15039.

[3] ZHAO Z X, XU S, ZHANG C X, et al. Bayesian fusion for infrared and visible images. Signal Processing, 2020, 177: 107734.

[4] SATHISH KUMAR G A E, DEVANNA H. Computational effective multimodal medical image fusion in NSCT domain. IOP Conference Series: Materials Science and Engineering, 2021, 1042(1): 012003.

[5] ZHANG H, XU H, TIAN X, et al. Image fusion meets deep learning: A survey and perspective. Information Fusion, 2021, 76: 323-336.

[6] ZHOU Q, YE S Z, WEN M W, et al. Multi-modal medical image fusion based on densely-connected high-resolution CNN and hybrid transformer. Neural Computing and Applications, 2022, 34(24): 21741-21761.

[7] GUO K, LI X F, HU X H, et al. Hahn-PCNN-CNN: an end-to-end multi-modal brain medical image fusion framework useful for clinical diagnosis. BMC Medical Imaging, 2021, 21(1): 111.

[8] SHI B B, CHEN Y N, ZHANG P, et al. Nonlinear feature transformation and deep fusion for Alzheimer's Disease staging analysis. Pattern Recognition, 2017, 63: 487-498.

[9] LIU S W, YANG L H. BPDGAN: a GAN-based unsupervised back project dense network for multi-modal

medical image fusion. Entropy, 2022, 24(12): 1823.
[10] XU H, MA J Y, JIANG J J, et al. U2Fusion: a unified unsupervised image fusion network. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.
[11] ZHANG H, MA J Y. SDNet: a versatile squeeze-and-decomposition network for real-time image fusion. International Journal of Computer Vision, 2021, 129(10): 2761-2785.
[12] LIU J Y, LIN R J, WU G Y, et al. CoCoNet: coupled contrastive learning network with multi-level feature ensemble for multi-modality image fusion. International Journal of Computer Vision, 2023, 132(5): 1748-1775.
[13] LIU X W, WANG R H, HUO H T, et al. An attention-guided and wavelet-constrained generative adversarial network for infrared and visible image fusion. Infrared Physics & Technology, 2023, 129: 104570.
[14] TANG L F, XIANG X Y, ZHANG H, et al. DIVFusion: Darkness-free infrared and visible image fusion. Information Fusion, 2023, 91: 477-493.
[15] VAN AARDT J. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. Journal of Applied Remote Sensing, 2008, 2(1): 1-28.
[16] LI S T, YANG B. Multifocus image fusion using region segmentation and spatial frequency. Image and Vision Computing, 2008, 26(7): 971-979.
[17] MANJUSHA D, BHOSALE U. Image fusion and image quality assessment of fused images. International Journal of Image Processing, 2010, 4(5): 484-508.
[18] QU G H, ZHANG D L, YAN P F. Information measure for performance of image fusion. Electronics Letters, 2002, 38(7): 313.
[19] ASLANTAS V, BENDES E. A new image quality metric for image fusion: The sum of the correlations of differences. AEU-International Journal of Electronics and Communications, 2015, 69(12): 1890-1896.
[20] ZHANG Y, XIANG W H, ZHANG S L, et al. Local extreme map guided multi-modal brain image fusion. Frontiers in Neuroscience, 2022, 16: 1055451.
[21] LI B, PENG H, LUO X H, et al. Medical image fusion method based on coupled neural P systems in nonsubsampled shearlet transform domain. International Journal of Neural Systems, 2021, 31(1): 2050050.
[22] LI B, PENG H, WANG J. A novel fusion method based on dynamic threshold neural P systems and nonsubsampled contourlet transform for multi-modality medical images. Signal Processing, 2021, 178: 107793.
[23] TANG W, HE F Z, LIU Y, et al. MATR: Multimodal medical image fusion via multiscale adaptive transformer. IEEE Transactions on Image Processing, 2022, 31: 5134-5149.
[24] SONG X, WU X J, LI H. A Medical Image fusion method based on MDLatLRRv2. arXiv preprint arXiv: 2206.15179, 2022.
[25] LIU Y, CHEN X, CHENG J, et al. A medical image fusion method based on convolutional neural networks//2017 20th International Conference on Information Fusion, July 10-13, 2017, Xi'an, China. New York: IEEE, 2017: 1-7.
[26] TAN W, THITØN W, XIANG P, et al. Multi-modal brain image fusion based on multi-level edge-preserving filtering. Biomedical Signal Processing and Control, 2021, 64: 102280.
[27] MA J Y, TANG L F, FAN F, et al. SwinFusion: cross-domain long-range learning for general image fusion *via* swin transformer. IEEE/CAA Journal of Automatica Sinica, 2022, 9(7): 1200-1217.

基于掩膜优化和并联注意力机制的多模态医学图像融合

邸 敬*, 梁 婵, 郭文庆, 廉 敬

兰州交通大学 电子与信息工程学院, 甘肃 兰州 730070

摘 要: 医学图像融合技术对于提高疾病的检测精确度和治疗效率至关重要,但现有的融合方法存在纹理细节模糊、对比度低,无法充分提取融合图像信息等问题。为此,提出了一种掩膜优化和级联式注意力机制的多模态医学图像融合方法。首先,将整幅图像转换为二值掩膜,构建轮廓特征图最大程度提升图像轮廓特征信息,并构建三重路径网络进行图像纹理细节特征提取和优化。其次,提出一种图像对比度增强模块和细节保留模块提升图像整体亮度和纹理细节。然后,利用通道特征和空间特征变化构造了一种并联注意力机制来融合图像,提升融合图像的显著信息。最后,设置了一个由残差网络构成的解耦网络,将融合图像与源图像之间的信息进行优化,旨在减少融合图像信息丢失。通过与近年来提出的9种高水平方法相比,本文方法的7项客观评价指标有6%—31%的提升,说明本文方法能够获得纹理细节更清晰、对比度更高和融合图像与源图像像素差异较小的融合结果,在主观和客观指标上都优于其他对比算法。

关键词: 多模态医学图像融合;二值掩膜;对比度增强模块;并联注意力机制;解耦网络

引用格式: DI Jing, LIANG Chan, GUO Wenqing, et al. Multimodal medical image fusion based on mask optimization and parallel attention mechanism. Journal of Measurement Science and Instrumentation, 2025, 16(1): 26-36. DOI: 10.62756/jmsi.1674-8042.2025003