High-Quality Fusion and Visualization for MR-PET Brain Tumor Images via Multi-Dimensional Features

Jinyu Wen, Asad Khan, Amei Chen, Weilong Peng, Meie Fang, C. L. Philip Chen, *Fellow, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—The fusion of magnetic resonance imaging and positron emission tomography can combine biological anatomical information and physiological metabolic information, which is of great significance for the clinical diagnosis and localization of lesions. In this paper, we propose a novel adaptive linear fusion method for multi-dimensional features of brain magnetic resonance and positron emission tomography images based on a convolutional neural network, termed as MdAFuse. First, in the feature extraction stage, three-dimensional feature extraction modules are constructed to extract coarse, fine, and multi-scale information features from the source image. Second, at the fusion stage, the affine mapping function of multi-dimensional features is established to maintain a constant geometric relationship between the features, which can effectively utilize structural information from a feature map to achieve a better reconstruction effect. Furthermore, our MdAFuse comprises a key feature visualization enhancement algorithm designed to observe the dynamic growth of brain lesions, which can facilitate the early diagnosis and treatment of brain tumors. Extensive experimental results demonstrate that our method is superior to existing fusion methods in terms of visual perception and nine kinds of objective image fusion metrics. Specifically, in the results of MR-PET fusion, the SSIM (Structural Similarity) and VIF (Visual Information Fidelity) metrics show improvements of 5.61% and 13.76%, respectively, compared to the current state-of-the-art algorithm. Our project is publicly available at: https://github.com/22385wjy/MdAFuse.**

***Index Terms*—MR, PET, multi-scale, image fusion, affine transformation.**

Manuscript received 25 September 2023; revised 02 April 2024; accepted 22 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62072126 and Grant 61772164, in part by the Fundamental Research Projects Jointly Funded by Guangzhou Council and Municipal Universities under Grant 202102010439, in part by the PolyU Research Institute for Sports Science and Technology under Grant P0044571, and in part by The Hong Kong Polytechnic University under Grant P0048387, Grant P0042740, Grant P0044520, Grant P0043906, Grant P0049586, and Grant P0050657. *(Corresponding author: Meie Fang.)*

Jinyu Wen, Asad Khan, Weilong Peng, and Meie Fang are with the Metaverse Research Institute, School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511400, China (e-mail: wjy1361120721@163.com; {asad, wlpeng, fme}@gzhu.edu.cn).

Amei Chen is with the Department of Radiology, Second Affiliated Hospital of South China University of Technology, Guangzhou 510641, China (e-mail: 17690189@qq.com).

C. L. Philip Chen is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China, and also with the Pazhou Lab, Guangzhou 510335, China (e-mail: philip.chen@ieee.org).

Ping Li is with the Department of Computing, the School of Design, and the Research Institute for Sports Science and Technology, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

I. INTRODUCTION

MEDICAL image fusion technology fuses important reference information from different modal images into one medical image. This resulting image provides more intuitive, comprehensive and clear information. Medical imaging plays an important role in various clinical applications. Among them, magnetic resonance (MR) and positron emission tomography (PET) provide an imaging basis for a variety of diseases and are widely used in the clinical diagnosis of diseases, such as benign and malignant brain tumors, depression, early-onset Alzheimer's disease, cerebral ischemia and others. The fusion of MR and PET brain images has been proven to be clinically significance [1], [2]. The diagnosis of a variety of brain diseases often depends on MR-PET imaging. Dynamic observation of brain lesions and qualitative analysis of brain tumors are also very important [3], [4]. MR imaging can clearly display soft tissue with high spatial resolution and is conducive to the determination of the scope of lesions. PET provides good physiological and metabolic information about the human body. Therefore, the extensive demand for combined MR-PET imaging has prompted the development of integrated MR-PET equipment (hybrid PET/MR). The ultimate effect of this advancement is to assist medical professionals in diagnosing brain abnormalities more effectively.

At present, the traditional methods for MR-PET medical image fusion are simple weighting [5], multiresolution pyramids [6], wavelet transform [7], color space [8], principal component analysis [9], human visual systems [10], etc. In the past decade, due to the emergence of large datasets and improvements in GPU computing power, the deep learning method, which does not rely on artificial features, has achieved great success in the field of image processing. Recently, deep learning has become a representative method of image fusion and a research hotspot. A variety of image fusion methods based on deep learning have been proposed in succession [11]–[13]. Fusing images with deep learning is an effective solution, to pursue a better perception effect, the design of an image fusion model is mainly based on the definition of some more diverse design rules to enhance the transformation and fusion strategy. In this paper, we propose a new idea for MR-PET medical image fusion based on unsupervised deep learning. First, a three-dimensional feature extraction module is established to extract the coarse, fine and multi-scale information features from a source image, and then the affine mapping function of multi-dimensional features is established to fuse the multi-dimensional features. In sum, our work makes the following three main contributions:

- Multi-dimensional analysis of the image is divided into three modules to extract the coarse features, fine features and multi-scale features from the source images. The three modules focus on the importance of different information from source images and try to extract different levels of features, reduce the loss of features in the process of feature transfer, and improve the accuracy of subsequent feature fusion.
- The affine mapping function is established to maintain the geometric relationship between different dimension features, whose correlation coefficients are generated adaptively through the learning process. So that multi-dimension features including the spatial texture information of the MRI image and the functional metabolic information of the PET/SPECT image can be fully preserved in the fused image.
- We propose a kind of energy-based color enhancement algorithm to further enhance the visualization effect, mainly enhance the energy information from the original PET/SPECT images in the fused image. While using it to display abnormal areas in different time series of the same case, the evolution process of lesions (e.g., brain tumors) can be tracked better.

The remainder of this article is organized as follows: the second section describes relevant work regarding medical image fusion. The third section demonstrates how to extract multi-dimensional features and fuse these features by a linear mapping function method. The fourth section describes the fusion experiment and result analysis of this method applied to MR-PET/SPECT brain images. The fifth section provides a summary of and futuristic prospective for the algorithm.

II. Related Work

Currently, the fusion methods that can be applied to MR-PET/SPECT medical images include some traditional fusion methods and deep learning based (DL-based) fusion methods, each type of which has its own advantages and disadvantages. In this section, we discuss some representative methods.

Traditional fusion methods: In traditional image fusion methods, the simple weighted averaging method, wavelet transform application and color space swap are three representative pixel-level fusion methods that are often applied to medical image fusion. Li and Wu [5] proposed a simple and effective image fusion method based on latent low rank representation (LatLrr) to better preserve the useful information in source images. This method uses a simple weighted average fusion strategy. By using the idea of low-rank clustering, the image is divided into low-rank and significant parts, and the weighted sum of the low-rank and significant parts is used to obtain the fusion result. Li et al. [14] used the low-rank decomposition method in noisy image fusion and obtained the fusion result mainly through minimum rank regularization. LatLrr can highlight the global structure information of an image, but its abilities for local structure preservation and detail extraction are poor.

Wavelet transform is also a classical method. Wavelets not only have orthogonality, biorthogonality and compactness but also have multiresolution characteristics. Zhan et al. [15] proposed an image fusion method based on phase congruency fusion (PCF); they used phase consistency to extract local and dramatic changes in images. This method uses a Gabor wavelet filter in space to improve phase consistency. The application of wavelet transforms [7], [16] is also conducive to understanding images, especially medical images. The wavelet transform has the characteristics of multiresolution and can observe signals from coarse to fine, but the wavelet transform method requires one appropriate mother wavelet and a feasible decomposition level. To facilitate doctors' understanding of images and especially to observe changes in physiological metabolism depicted in PET images, some researchers have developed fusion strategies to maintain pseudocolor. Du et al. [8] used a dual-scale strategy and color domain transformation to maintain pseudocolor information, thereby integrating gray images (such as MR) and pseudocolor images (such as PET and SPECT), and focused on preserving pseudocolor information. However, color space conversion will also cause some information loss. In conclusion, some traditional fusion methods can achieve a high-quality fusion effect, but most fusion methods depend on manual feature extraction rules of specific image types, including setting parameters. With increasing image type and number, feature extraction becomes increasingly complex. At the same time, the generalization ability of traditional methods is very weak.

DL-based fusion methods: In multimodal image fusion, traditional methods and deep learning methods are combined, which can improve model performance. Zhong et al. [17] proposed a joint image fusion and superresolution method based on a CNN. Rajalingam et al. [18] proposed a deep guided hybrid multimode medical image fusion (HMMIF) method that has been applied to the neuropathology of neurocysticercosis, a degenerative disease. The combination of classic traditional and deep learning methods can effectively improve the performance of networks. However, some traditional methods need a priori knowledge, and combining traditional methods may increase the time and space complexity. Recently, some end-to-end deep neural networks have been applied to medical image fusion [19], [20]. Rajalingam and Priya [21] proposed a multimodal medical image fusion method based on a deep learning neural network. The method used a conjoint convolutional neural network to generate a weighted graph to fuse pixel motion information from multimodal medical images. Zhao and Zhao [22] proposed a general fusion framework based on representation learning, which study is a domain-specific unreferenced perceptual metric loss based on edge detail and contrast to optimize the learning process and make the fused images exhibit a more specific appearance.

In order to improve the global feature coding capability of U-Net, Xiao et al. [23] introduced global feature Pyramid extraction module (GFPE) and global attention connection on sample module (GACU) to extract and utilize global semantic and edge information effectively. Unsupervised deep neural network models such as DeepVTF [24] and VIFNet [25] have been designed recently. The DeepVTF method established a

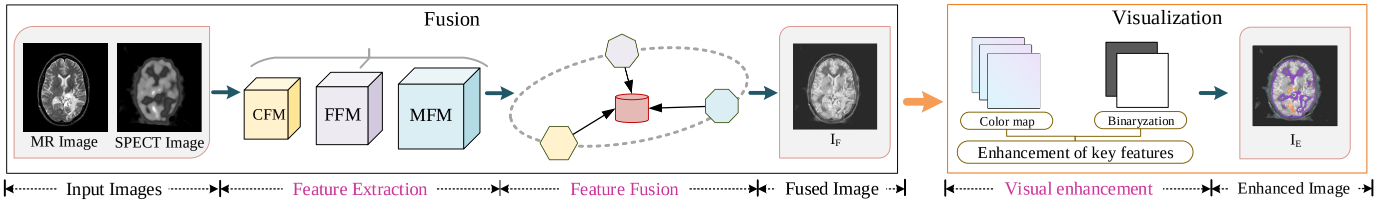


Fig. 1. The flowchart of MdAFuse consists of two stages, i.e., image fusion and visualization enhancement. A novel CNN network is built to fuse a pair of MR-SPECT image to obtain a piece of high-quality fused image in the fusion stage. And in the visualization stage, a new color enhancement algorithm is proposed to further improve the visualization effect of the fused image. More details for these two stages are respectively displayed in Fig. 2 and Fig. 3.

visual similarity measure between the input conventional true color and the fused output to obtain a natural and intuitive image. VIFNet is derived from a robust hybrid loss function, which is composed of a modified structural similarity measure and total variation. Kumar et al. [26] proposed a deep neural network model called FunFuseAn, which uses SSIM as the loss function to fuse MRI and PET images. Xu et al. [27] introduced a depth fusion model based on gradient and connected domain multifocus images. To overcome the obstacle of gradient loss when a deep network was applied to multifocus image fusion, a mask network was designed to generate a binary mask directly. The consistency verification strategy was adopted to generate fusion results by adjusting the initial binary mask. Zhang et al. [28] proposed the fusion model IFCNN, which uses transfer learning technology and maximum tensor strategy. Xu et al. [29] introduced an unsupervised and unified dense connected network called FusionDN for different types of image fusion tasks. This method generates a fused image based on the degree of retention of the source image (i.e., the quality of the source and the amount of information) by training dense connected networks.

Specially in recent two years, several new methods have been proposed, as [30]–[32]. Zhao et al. [30] proposed a dual-branch Transformer-CNN network architecture (CDDFuse) for multimodal image fusion. To extract specific modality features and modality-shared features, they employed Restormer, Lite Transformer, and reversible neural network modules. Feature decomposition was achieved through the introduction of a correlation-driven decomposition loss. Xu et al. [31] proposed a novel unified unsupervised end-to-end image fusion network (U2Fusion). Through feature extraction and information measurement, U2Fusion automatically estimates the importance of correspondences from source images and provides adaptive information retention. Liang et al. [32] proposed a powerful image decomposition model called Decomposition for Fusion (DeFusion) that performs fusion tasks through self-supervised representation learning without any paired data or complex loss functions. DeFusion can decompose the source images into a feature embedding space where common and unique features can be separated, allowing image fusion to be achieved within the embedding space through reconstruction jointly trained during the decomposition stage, even without any fine-tuning. It designs a self-supervised pre-training task based on common and unique decomposition (CUD) that adapts to image fusion task. Another important branch of the deep learning method is a method based on a generative adversarial network (GAN) [33], [34]. Inspired by the conditional generative adversarial network (CGAN), Ma et al. [35] introduced a GAN into image fusion and an unsupervised GAN network image fusion framework termed FusionGAN.

These newly-developed deep learning models have designed a unified framework for image fusion, which is powerful and performs well for fusing natural images. They can also be directly used for medical image fusion with better performance. Compared with the traditional methods, the deep learning method has the characteristics of independent learning, flexible real-time processing and good generalization. Considerable potential has been identified in deep learning-based methods. However, there are significant differences between medical images and natural images, and the dataset size is small, resulting in a general decline in the performance of fusion models designed for natural images on medical images. At present, deep neural networks specifically designed for medical image fusion are still in their infancy. Corresponding methods that can be applied in hybrid PET/MR are even more rare. In order to achieve better model performance with limited data samples, medical image fusion typically requires consideration of specific clinical insights and imaging mechanisms (such as mechanisms and anatomical structures of different modalities), and the design of effective fusion strategies to integrate these different information sources, thereby fully utilizing these medical domain knowledge in the fusion process. In this paper, we focus on extracting and preserving key features from MR and PET images. To achieve this, multi-dimensional feature models and an adaptive linear fusion strategy are designed.

III. The Proposed Image Fusion Network and Visualization Method

As illustrated in Fig. 1, our work in this paper consists of two aspects. Firstly, we propose a DL-based fusion network designed for MR and PET/SPECT brain images. Secondly, we present an energy-based visualization method to further enhance the fused images. In the fusion network, we employ a multi-dimensional feature extraction method and an adaptive linear fusion strategy. These aspects will be discussed in detail in Subsection A and Subsection B, respectively. The Subsection C will cover the explanation of how the loss function of this network is set. Subsequently, we will describe the proposed visualization method in Subsection D.

A. Feature extraction

In this procedure, we try to extract and preserve important information from the source MR and PET images as much

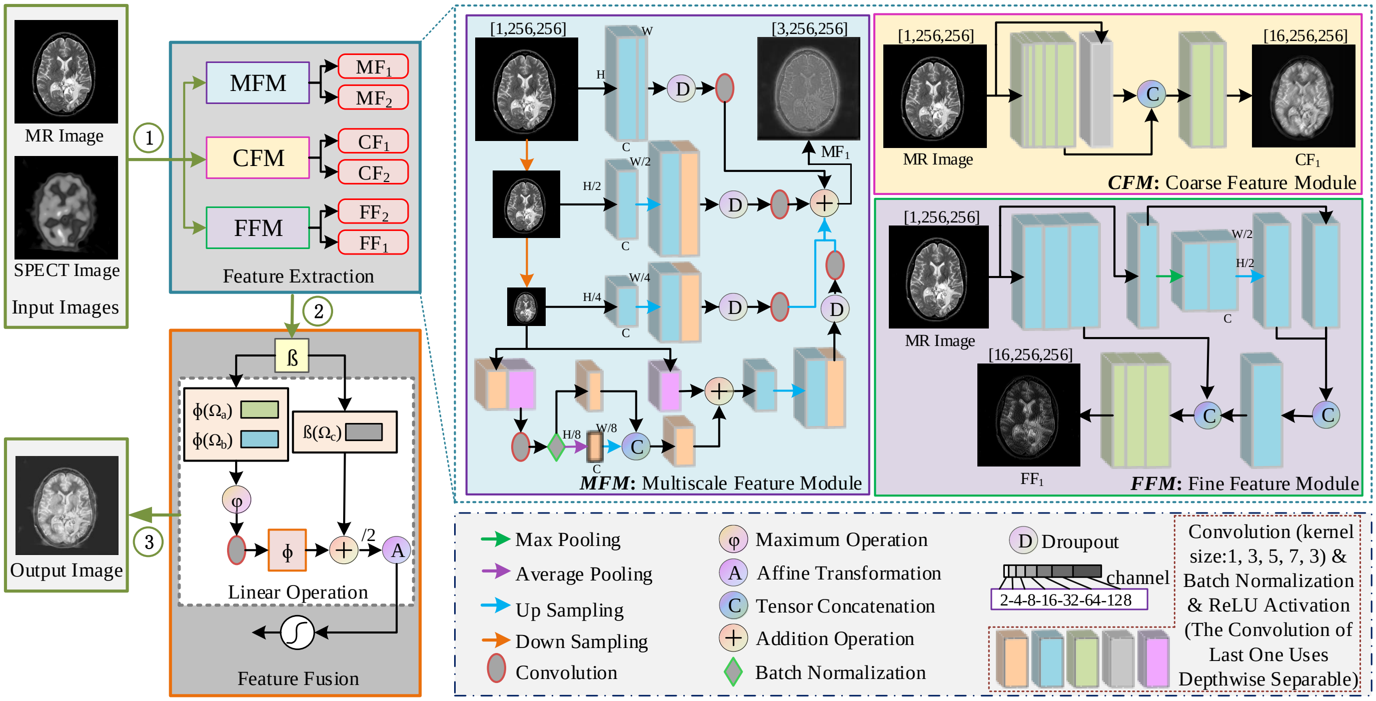


Fig. 2. The framework of the proposed image fusion network, including feature extraction and feature fusion. The feature extraction consists of three modules: CFM (coarse feature module), FFM (fine feature module) and MFM (multi-scale feature module). The feature fusion adopts a adaptive linear mapping function to establish relationship between multi-dimensional features.

as possible, we adopt a multi-dimensional approach, making full use of spatial and channel features, and establish three modules, namely, the coarse feature module (CFM), the fine feature module (FFM) and the multi-scale feature module (MFM), to extract coarse feature information, fine feature information and multi-scale feature information, respectively. CFM is responsible for capturing the main shape structure and edge contour information of images, which belongs to key information for distinguishing tissue boundaries and lesion localization. FFM is responsible for capturing fine texture details, which belongs to key information for identifying lesion types. MFM is mainly designed to retrieve the complementary information between multi-scale images to achieve finer fusion.

As illustrated in Fig. 2, the specific feature extraction process is presented in the blue dotted line box. In Fig. 2, the small squares with different colors represent three steps (convolution operation, batch normalization and ReLU activation), different colored squares represent different sizes of convolution kernels, yellow is 1, blue is 3, green is 5, purple is 7 and pink is 3. Squares of different sizes represent different feature map sizes and the number of output channels. There is a definition of different channel numbers in the gray dotted box at the lower right corner of Fig. 2, that is, squares with different widths represent the outputs of different channel numbers, and the darker the color is, the bigger the number of channels. For an input image, the feature extraction module will obtain three dimensions of features, namely, CF_1, FF_1 and MF_1. These feature extraction modules are designed based on the perceptual characteristics of human visual perception and the physical properties of images.

In Fig. 2, CFM is represented by the area with a yellow background on the edge of the fuchsia solid line. CFM stands for coarse feature module, which is responsible for coarse feature extraction from images through convolution operations. It utilizes larger convolution kernels and smaller feature channels to capture the main shape structures and edge contour information of images, representing global low-frequency information such as image backgrounds and object contours. From the feature maps shown in Fig. 2, it can be observed that the CF feature map extracted by the CFM module exhibits clearer contours. Similarly, in Fig. 2, FFM is denoted by the area with a purple background on the edge of the blue solid line. FFM refers to the fine feature module, which focuses on fine feature extraction. This module employs smaller convolution kernels and larger feature channels to capture texture details such as object edges, fine texture details, and noise, representing local high-frequency information with significant variations within the images. The number of channels used for fine feature extraction is much greater than that used for coarse feature extraction. From the feature maps shown in Fig. 2, it can be observed that the FF feature map extracted by the FFM module displays more distinct texture details.

In Fig. 2, MFM is the area with blue background on the edge of purple. MFM mainly obtains the spatial structure from the idea of pyramid and classification to obtain the complementary information between multi-scale images and uses the complementary information between multiresolution and multi-scale images to achieve fine fusion. In the MFM, the deep separation convolution network is also used, which is convenient for mining more abundant and useful information and can provide a good foundation for subsequent image comprehension and

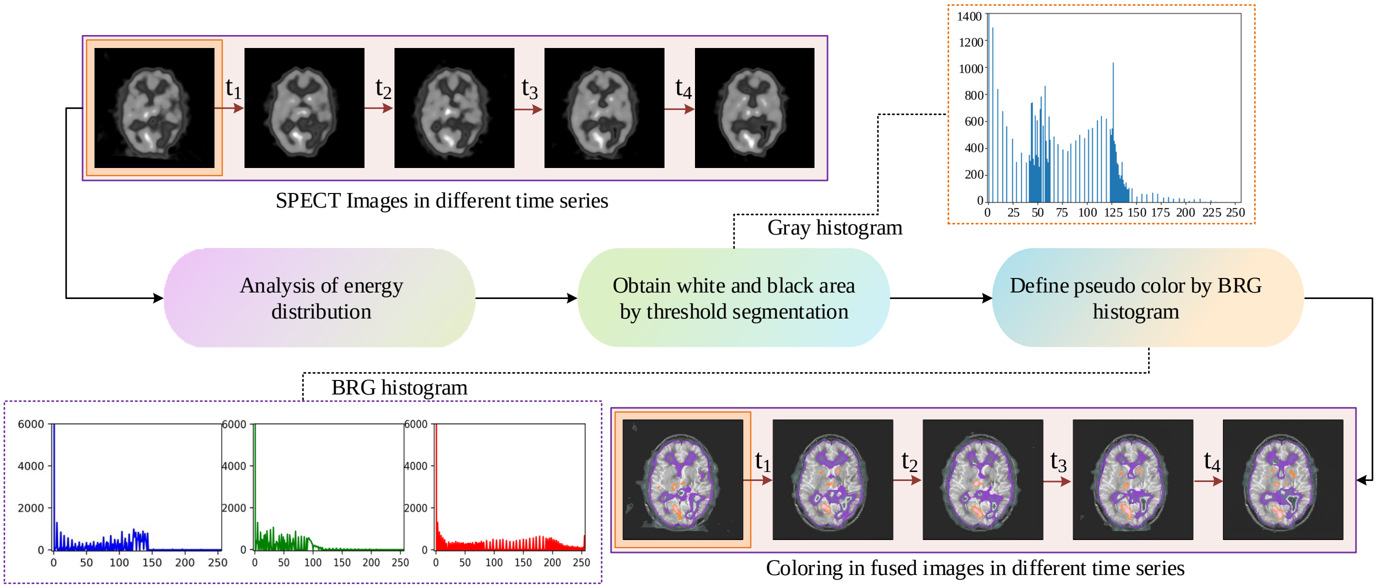

t1
t2
t3
t4
SPECT Images in different time series
Gray histogram
1400
1200
1000
800
600
400
200
0
0 25 50 75 100 125 150 175 200 225 250
Analysis of energy distribution
Obtain white and black area by threshold segmentation
Define pseudo color by BRG histogram
BRG histogram
6000
4000
2000
0
0 50 100 150 200 250
Coloring in fused images in different time series

Fig. 3. The pipeline of defining pseudo color by BGR histogram to enhance the energy information of original SPECT image in the fused image. Given five pieces of SPECT images captured at different time series from the same case, this pipeline generates the corresponding enhanced fused image series, which facilitate dynamically tracing abnormal regions and diagnosing brain tumors.

application analysis. As in reference [36], [37], we use a skip connection approach when implementing neural networks. In our design of the feature extraction module, we place more emphasis on local fine-grained information rather than global features. Using short skip connections facilitates the transmission and preservation of local information, enabling the network to better learn and utilize these localized characteristics. Therefore, short skip cascading is used in the process of feature extraction to obtain more detailed information, realize the gradual fusion of features.

B. Fusion strategy

MdAFuse extracts features from three dimensions, i.e., coarse, fine and multi-scale dimensions. The coarse features are mainly reflected in the edge contour information, while the fine features can retain the texture details. To obtain both high frequency information and low frequency information from an image, the coarse and fine features of each image are added to obtain the image features with more information. To highlight salient features in two images, especially highlighted information on PET/SPECT images (often corresponding to abnormal areas), we use the maximum value to obtain the more prominent part. Multi-scale features can provide more structural information and can also compensate for missing local details at the cross scale. Therefore, the coarsest and finest features must be multi-scale. For each input, three features (CF, FF and MF) will be obtained. When two modal images are input, six features are extracted.

We design an affine transformation function to fuse them together. According to the reference [38], affine transformations exhibit unique geometric fidelity properties that demonstrate distinct advantages within the context of linear fusion frameworks. Further, we set the affine transformation scale to learnable parameters to build an adaptive fusion strategy, which can effectively enhance both the overall quality of fused images and the integrity of information they contain. So an affine transformation is adopted as an adaptive strategy to make full use of the correlations among features and maintain them as:

$$\Phi(\Gamma) = \beta(\sum_{i=1}^{U} CF(i))+\beta(\sum_{i=1}^{U} FF(i)) = A\sum_{j=1}^{2}\sum_{i=1}^{U}\Gamma_j(i)+2e, \tag{1}$$

where Φ is the sum of two features that obtained by affine transformations, as shown in Eq. (1). When Γ represents the elements or properties of Ω_a, Γ_j denotes a subset or component of Γ that encapsulates both the CF_1 feature and the FF_1 feature. In Eq. (1), U means the size of feature map. $\beta(\Gamma) = A\Gamma(i) + e$, $A = cI$, where I is the identity matrix, c is the scaling factor and e represents the shift vector. The proposed fusion strategy is adaptive because both c and e are learnable parameters. The rules of setting the size are the same as those for post-normalization [39]. The value of initialization c is 1, and e is 0. β is the affine transformation function, which is used to express the correlation between features. The affine transformation is geometrically defined as a nonsingular linear transformation and translation transformation between two vector spaces, which can be used to maintain the position relationship between two-dimensional images. We adopt the emulating objective function as:

$$F = \delta(\Phi((\upsilon(\varphi(\Phi(\Omega_a),\ \Phi(\Omega_b))) + \beta(\Omega_c))/2)), \tag{2}$$

where F is the result of fusion. δ is the activation function, $tanh$ being adopted in this paper. $\Omega_a = \{CF_1\} \cup \{FF_1\}$, $\Omega_b = \{CF_2\} \cup \{FF_2\}$, and $\Omega_c = (MF_1 + MF_2)/2$. υ represents a convolution operation to keep the input and output dimensions consistent. φ is the maximum operation to highlight the significant information.

C. Loss function

To reconstruct the source image more accurately, we consider the feature extraction stage and fusion process in the training phase and obtain the minimum loss function:

$$L = \alpha L_S + \lambda L_M. \tag{3}$$

The total loss function consists of structural similarity L_S and pixel loss L_M, The specific calculation process is provided in the Appendix. Here, α, λ are weight parameters for users to set in the interval $[0, 1]$. Through comparative analysis, α set to 0.3 and λ set to 0.7 in the training process, and the result is the best. In the training process, the initial learning rate is 0.0001, 60 epoch, batch size is 2, we employ batch normalization, and the optimization function is Adam.

D. Enhancement of key features

PET or SPECT image has limited spatial resolution and unclear grayscale distinctions at boundaries, constraining the differentiation between abnormal and normal areas. To enhance visibility, pseudocolor processing is commonly applied in clinical settings, assisting doctors in identifying boundary details and improving the accuracy of disease diagnosis.

In the generation of grayscale fused images through MR-PET/SPECT fusion, spatial textures from the MR images are clear, while the low-resolution grayscale PET/SPECT images struggle to reveal significant abnormalities in glucose metabolism. Existing pseudocolor methods involve discrete color mapping based on grayscale and continuous translation from grayscale to color properties. However, none of these techniques focuses on highlighting these significant abnormalities in PET/SPECT images. To emphasize these abnormalities, we propose an energy-based color enhancement algorithm based on the grayscale histogram of PET/SPECT. This algorithm enhances the visual effect of abnormal information, and details are described in Algorithm 1 and Fig. 3. where, I_g represents the gray PET/SPECT, I_f is the fused image, I_e is the enhanced image, η_w and η_b are the white and black areas in the threshold binarization process respectively. ρ_w and ρ_b are two self-defined color for the white and black areas. $I_c[a_{ij}](i,\ j)$ represents the region in the I_c correspingding to I_a. The key feature enhancement process is shown in Fig. 3. In this part of the work, the main purpose is to obtain useful information that can represent the glucose content or physiological metabolism in a PET/SPECT image and enhance important information in fusion results as:

$$\zeta = \vartheta\left(\chi\left(\phi\left(I_g\right)\right),\ p\right), \tag{4}$$

$$\tau_w = \zeta^l,\ l = \kappa - 1, \tag{5}$$

$$I_w(i,\ j) = \begin{cases} 0, & I_g(i,\ j) \leq \tau_w \\ 255, & I_g(i,\ j) > \tau_w \end{cases}, \tag{6}$$

$$I_b(i,\ j) = \begin{cases} 0, & I_g(i,\ j) < \tau_{b1} \parallel I_g(i,\ j) > \tau_{b2} \\ 255, & I_g(i,\ j) > \tau_{b1}\ \&\ I_g(i,\ j) \leq \tau_{b2} \end{cases}. \tag{7}$$

Algorithm 1. Energy-Based Color Enhancement

```
Input: Images $I_g$, $I_f$
Output: Image $I_e$
$x \leftarrow$ the weight of $I_g$;
$y \leftarrow$ the height of $I_g$;
$\rho_w \leftarrow$ self-defined color for image of white area;
$\rho_b \leftarrow$ self-defined color for image of black area;
$\eta_w,\ \eta_b \leftarrow \gamma(I_g)$;
$\sigma \leftarrow$ ColorUp($I_g,\ I_f,\ \eta_w,\ \rho_w$);
$I_e \leftarrow$ ColorUp($I_g,\ \sigma,\ \eta_b,\ \rho_b$);
____________________________________________;
Function ColorUp($I_g,\ I_f,\ I_a,\ I_c$)
    $I \leftarrow I_f$.copy();
    if $I_c == \rho_w$ then
        $\mu \leftarrow 5\sigma - 20,\ \sigma = 0, 2, 6, 8$;
        $\xi \leftarrow 5\iota - 20,\ 0 \leq \iota \leq 4$;
    else
        $\mu \leftarrow 5\sigma - 20,\ 0 \leq \sigma \leq 11\ \&\ \sigma \neq 4$;
        $\xi \leftarrow 5\iota + 10,\ 0 \leq \iota \leq 11$;
    $n \leftarrow len(\mu)$;
    for $i$=1 to $x$ do
        for $j$=1 to $y$ do
            if $I_a(i,\ j) \neq 0$ then
                $a_{ij} \leftarrow I_a(i,\ j)$;
                $\psi \leftarrow I_f(i,\ j) - I_g(i,\ j)$;
            for $k$=1 to $n$ do
                $I^{'} \leftarrow I[a_{ij}]$;
                $I_c^{'} \leftarrow I_c[a_{ij}]$;
                if $\psi \leq \mu[k]$ then
                    $I^{'}(i,\ j) \leftarrow I_c^{'}(i,\ j) + \xi[k]$;
                else
                    $I^{'}(i,\ j) \leftarrow I_c^{'}(i,\ j) + \xi[-1]$;
    return $I^{'}$;
```

Firstly, we should analyze the energy distribution in the gray PET/SPECT, including the size distribution of different pixels, distribution of different positions of pixels, concentration areas of useful information, and number of regions representing similar information. There are two forms of energy distribution mainly in PET/SPECT. One is the tissue with high glucose content or vigorous metabolism, which is displayed in white area. The other is the tissue with low glucose content or weak metabolism, which is shown in black area.

Secondly, we need to obtain black and white areas in gray PET/SPECT. We use the thresholding method to extract these two types of images. The key to thresholding lies in the setting of the threshold, and τ_w is the threshold of extracting the white area block, which is obtained by Eq. (5). In Eq. (5), ζ refers to the grayscale values of the peaks in the histogram. κ refers to the total number of peaks. In Eq. (4), the function of ϕ represents the computation of the histogram, the function of χ represents its transformation into one dimension, and the parameter p denotes the minimum height difference between the peak and its surrounding valleys, which set 100 in the example of Fig. 3. The function of ϑ is used for finding the peaks in the histogram. For the black area, we set two thresholds. To eliminate some noise and avoid false overflow, we add another threshold τ_{b1}, which is set to 10. τ_{b2} is the global threshold obtained by the Otsu method. γ in Algorithm

1 means the computation of two binary images, which are the threshold binary images of the white feature and black feature obtained by Eq. (6) and Eq. (7), respectively. In Eq. (6) and Eq. (7), $I_g(i,\ j)$ represents the pixel value at the position of the i-th row and j-th column in image I_g.

Finally, we DeFusionine the pseudocolor according to the BGR histogram, the specific process as shown in Algorithm 1, and three images in Fig. 3 are included in the histogram. The histogram of the three colors in the corresponding BGR, which calculated from the pseudocolored PET/SPECT. For the convenience of observation, the value of the vertical axis is only less than 6000. From the BGR histogram, we can determine the approximate range of blue, green and red pixel values. Therefore, the white area assumes an yellow display, and the black area assumes a dark blue display. ρ_w and ρ_b used to represent yellow and dark blue area respectively. The brightness varies according to the size of pixels and does not display colors for areas that do not have auxiliary diagnostic value. Then, the pixel value between the fused image and the original PET/SPECT gray image is compared, and corresponding changes are made. The specific process as shown in the function of $ColorUp$ of Algorithm 1, and the coloring effect of the resulting image is illustrated in Fig. 3.

IV. Experimental Results

We used Python 3.7.9 and the Pytorch version of 1.7.1 + cu110 with its corresponding CUDA version of 11.1. GNU/Linux x86 in a GeForce RTX 3090 Ti GPU 20 GB RAM device-64 system. In the training, the data came from the Alzheimer's Disease Neuroimaging Initiative unites researchers (https://adni.loni.usc.edu/data-samples/access-data/), where we obtained 555 pairs of MR and PET images. The age of these samples ranges from 55 to 90 years, and the sex includes both male and female. All images were analyzed as axial slices with a voxel size of $1.0\times1.0\times1.0\ mm^3$. Cross-validation are performed for our trained model by non-intersect test dataset. 137 pairs of medical images from the Whole Brain Atlas (http://www.med.harvard.edu/AANLIB/home.html.) are obtained for testing. Among them, there are 74 pairs of MR-SPECT, 42 pairs of MR-PET, and 21 pairs of MR-CT.

In the fusion experiment of Fig. 4 and Fig. 6, the test data were MR and SPECT brain images with glioma, which is a typical application in the brain nervous system. They are identified in the test dataset and respectively correspond to No. 010 and No. 029. Fusion of MRI and PET/SPECT can integrate biological anatomical information and physiological metabolic information, which can help doctors to locate and diagnose lesions. Several existing fusion algorithms based on traditional and deep learning methods are analyzed and compared.

In order to quantitatively measure the results obtained from different fusion methods, we adopted nine kinds of quality evaluation metrics, i.e., information entropy (EN) [40], mutual information (MI) [41], standard deviation (SD) [42], visual information fidelity (VIF) [43], structural similarity (SSIM) [44], sum of correlations of differences (SCD) [45], correlation coefficient (CC) [46], mean square error (MSE) [47] and the ratio of spatial frequency error (rSFe) [48]. For MSE, a smaller value signifies better performance. For rSFe, a smaller absolute value of rSFe corresponds to a better fusion effect. While other seven evaluation metrics are on the contrary. Among them, rSFe is a relatively uncommon evaluation metric, which is used to measure the overall activity level of the image, which consists of four spatial frequencies (row, column, main diagonal, secondary diagonal) and four first order gradients (horizontal, vertical, main diagonal, secondary diagonal) in pixel points. The smaller the absolute value of this index is, the better the fusion effect is. The specific calculation equation of rSFe is explained in the Appendix.

TABLE I
Linear transformations on each metric, where k_i and b_i represents the transformation coefficients in Fig. 5, Fig. 7, and Fig. 10, respectively.

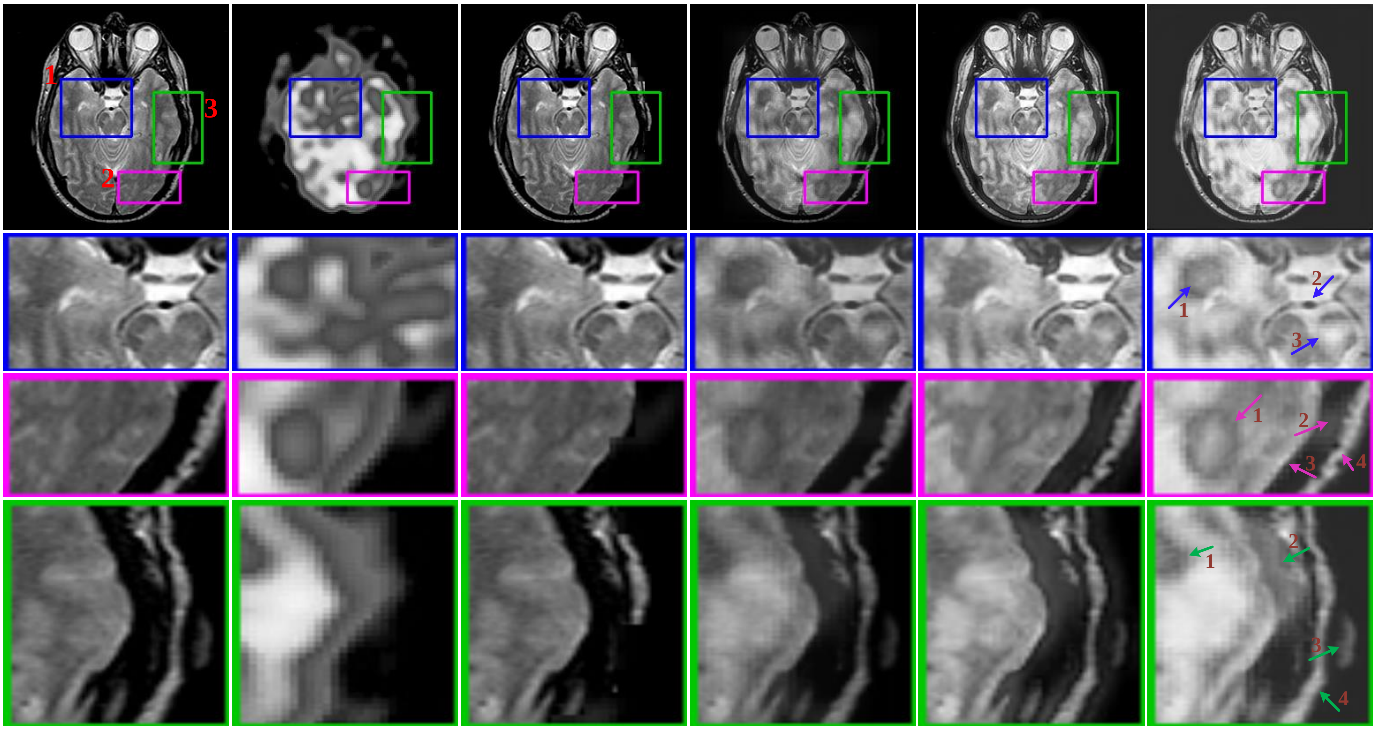
	EN	MI	SD	VIF	SCD	SSIM	CC	MSE	rSFe
k_1	8	6	1/2.1	48	29	16	42	18	-37
b_1	-11	0	0	0	2	0	5	5	8
k_2	5.6	5	0.4	31	18	16	26	240	-32
b_2	0	0	0	0	0	0	0	3	6
k_3	1	1	1/13	5	2	10	100	15	-8
b_3	-1	0	0	0	0	-14	90	0	0

For the convenience of comparison and analysis, we normalized the calculation results of some metrics. As shown in the Table I, we performed linear transformations on each metric, where k_i and b_i $(i = 1, 2, 3)$ represent the transformation coefficients in Fig. 5, Fig. 7, and Fig. 10, respectively.

A. Comparisons with traditional fusion methods

In this section, we test three kinds of traditional fusion methods and compare them with the method proposed in this paper. As illustrated in Fig. 4, the first column and second column display the MR and SPECT source images, respectively, from some cases with brain abnormalities. The following columns show the results obtained by three traditional methods and MdAFuse, with one column for each method. In Fig. 4, we can observe the differences among the fused images from different fusion methods. PCF is a fusion method using wavelet transform, and the fused image has serious information loss. LatLrr is a low-rank representation method, and its fusion strategy is simple weighted average. The results of this method can maintain good texture details but less SPECT image information. atsIF (adaptive dual scale image fusion) [8] is a fusion strategy through color space conversion, which can retain more SPECT information, but MR details are lost. Compared with other methods, MdAFuse can better preserve the information of the two source images. In the first line, three corresponding parts of each image inside three colorful squares are enlarged and shown in the 2nd, 3rd and 4th lines for clearer comparison.

The second line in Fig. 4 lists the enlarged images corresponding to the parts inside blue squares in the first line. In the last column, we mark three arrows pointing to the temporal lobe, basilar artery and pontine. The temporal lobe is gray and irregular in the MR image. SPECT shows a lighter gray circle with an obvious color difference. In the PCF results, the area is not obvious, and the features are completely retained on MR



(a) MR Image (b) SPECT Image (c) PCF (d) LatLrr (e) atsIF (f) MdAFuse

Fig. 4. Fused results and their enlarged ROIs of the same pair of MR-SPECT images by some traditional fusion methods (PCF [15], LatLrr [5], atsIF [8]) and the proposed fusion method.

but not on SPECT. In the results of LatLrr fusion, the area can retain the information of the two source images but also loses part of the SPECT information, such as the pixel value around the temporal lobe, which is not different. The effect of the atsIF method is greatly improved in terms of brightness, which can retain more information in SPECT, but information loss occurs with the combined source image. Compared with other methods, MdAFuse can perfectly preserve the features of the temporal lobe region. For the basilar artery area, an obvious difference is also noted. In the MR image, the area is a small black ellipse, while in SPECT, it is an even gray area without shape features. PCF completely retains the information on MRI, LatLrr and MdAFuse show a dark gray color for this position area, and the shape is not changed. The fusion result of atsIF becomes a white area, and a black circle is evident outside the small ellipse. Arrow 3 points to the pontine area, which is an area with an uneven gray distribution on MR images and a small area with obvious brightness similar to triangle on SPECT images. Similarly, the feature information of SPECT cannot be found in the fusion results of PCF, and the brightness information of this location area is not obvious in the LatLrr and atsIF results. The features of MR and SPECT can be clearly viewed in the results of this method.

The third line in Fig. 4 lists the enlarged images corresponding to the parts inside purple squares in the first line. Similarly, in the last column, we mark four arrows pointing to the occipital lobe, occipital bone, lateral sinus and occipital lobe contour. From the display of the four arrows, it also proves that the fusion result of MdAFuse is better especially in the area of first arrow. The fourth line in Fig. 4 lists the enlarged images corresponding to the parts inside blue squares in the first line. The four arrows in the last column point to the temporal lobe, temporal muscle, pinna and lateral occipital bone. From the display of the four arrows, it also proves that the fusion result of MdAFuse is better.

Traditional image fusion methods mainly belong to pixel-level operations in which pixel-level fusion can be regarded as the information only for feature extraction and direct use. This kind of method pursues the maximum amount of information to retain. To fully prove the effectiveness of the proposed image fusion model, Nine metrics (EN, MI, SD, VIF, SSIM, SCD, CC, MSE, rSFe) are used to evaluate the quality of the fusion results which shown in the Fig. 5. In Fig. 5, it can be observed that most of the metrics for our MdAFuse method hold advantages. Specifically, the SCD value is 9% higher than the sub-optimal value.

B. *Comparisons with deep learning fusion methods*

In this section, we conduct experiments on several kinds of DL-based image fusion methods and compare their fusion results. As illustrated in Fig. 6, the first and second columns show the MR and SPECT source images to be fused, and the following six columns show the fusion results obtained by six kinds of fusion methods, including MdAFuse. From these results, we can observe the differences among different fusion methods. Overall, it can be observed that the fusion images generated by U2Fusion and CDDFused seem not to consider the information of SPECT. While FusionDN and DeFusion methods preserve SPECT information in the fusion images, some loss of information can still be observed. The

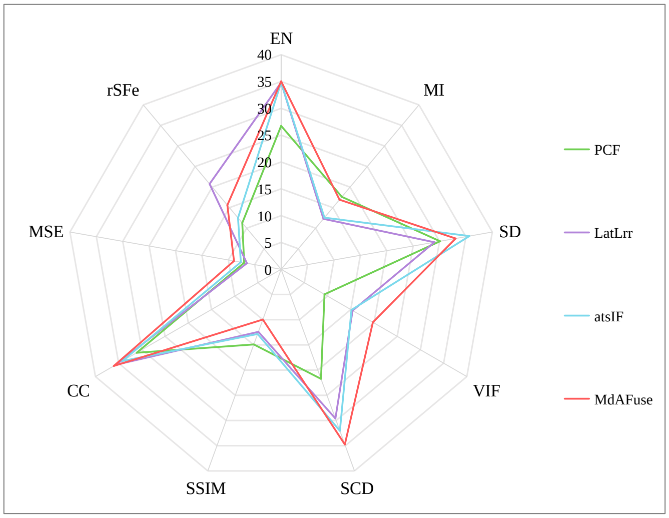


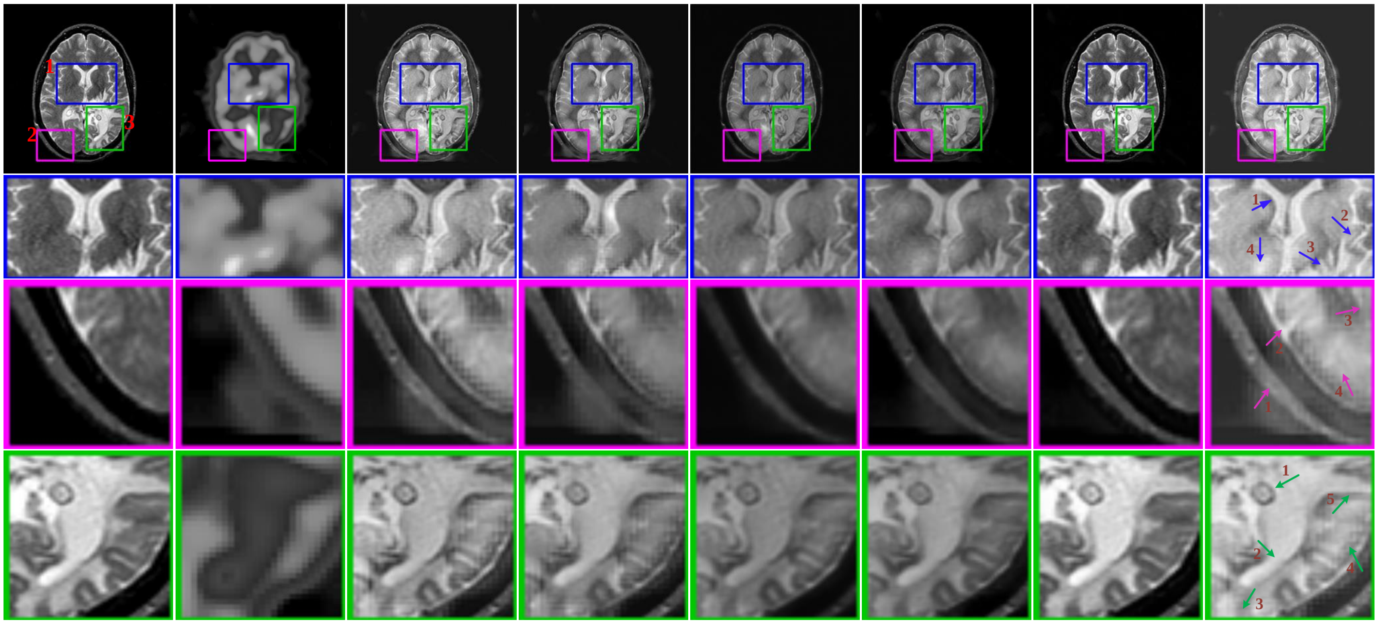
Fig. 5. Evaluation metrics of different fusion methods on MR-SPECT image pairs corresponding to Fig. 4.

image fusion results from the FunFuseAn method are relatively better, showing visual similarity to MdAFuse in this paper. Therefore, we enlarged three small areas to further compare and analyze. In the first line, we use three colors to mark three different areas and enlarge them, as shown in the second to fourth lines.

Firstly, the ROIs marked with the blue square in Fig. 6 mainly reflect the information of the frontal horn of the lateral ventricle, thalamus and temporal lobe. In MR images, the frontal horn of the lateral ventricle is white and symmetrical, while in SPECT images, the left lateral ventricle is gray. The area of the left thalamus is dark gray in MR images, and the lenticular area is bright in SPECT images. The area of the right thalamus is white and looks like a jungle in MR images, while in SPECT images, it is black and irregular. The area of the temporal lobe is measured on the left and right sides of the blue square, which is mainly reflected in the MR image and is continuous. In the last column of Fig. 6, four areas are marked with colorful squares. Area 1 is the left frontal horn of the lateral ventricle. The structure of this position in U2Fusion and DeFusion is relatively complete, but the image looks fuzzy. CDDFuse highlights the MR information. FunFuseAn is similar to FusionDN, and both have color aberrations. FusionDN of the left frontal horn of the lateral ventricle is complete, but the pixel value distribution of right frontal horn of the lateral ventricle is uneven. In contrast, MdAFuse can preserve the MR features more completely. Area 2 is the temporal lobe. In the results of FunFuseAn and FusionDN, the structure is relatively complete and continuous, but the color is relatively dark. The results of U2Fusion and DeFusion can well preserve the MR features, but artifacts exist. Area 3 is the location to the right of the thalamus, which should reflect more texture details on MR images. In the result of U2Fusion and DeFusion, the location is not obvious. FunFuseAn, FusionDN and CDDFuse are similar because the organizational structure is incomplete. In contrast, MdAFuse can preserve the right thalamus completely. Area 4 is the left thalamus, which should show a diffusion area of a bright spot. The results of FusionDN and DeFusion show no obvious state. The results of U2Fusion can identify a brightness block, but the grayscale value of this area is more even, and no difference change is found. Other methods can find the brightness block obviously and can detect more layers of brightness features.

Secondly, we noticed that in the enlarged second purple square in Fig. 6, the four areas marked in the last column are the meninges, occipital lobe, occipital horn of lateral ventricle and cerebral cortex (gray matter). In the MR image, area 1 appears as a thick gray column, while the SPECT image shows an uneven wavy gray block. The results of FusionDN and DeFusion all show some faults, and the complete structural information of MR is not preserved. U2Fusion and CDDFuse mainly retains MR information and ignores the SPECT characteristics. In contrast, FunFuseAn and MdAFuse can consider the more important information of the two original images and retain the useful features more completely. Area 2 is the occipital lobe, which is not clear to find in the U2Fusion and DeFusion results. No significant difference is noted between the location area and the surrounding area in the results of FusionDN. Area 3 is the left occipital horn of the lateral ventricle. We can find the characteristic information of the occipital horn on the MR image, which is an area with even grayscale values on the SPECT image. FunFuseAn can find subtle changes, but the result of our method is more obvious, and the structures with other methods in this position are not very clear. Arrow 4 refers to part of the cerebral cortex (gray matter). In this area of the MR image, we can find a scattered distribution of uneven grayscale values, and we can observe the fine texture features. In the SPECT image, the area is light gray with even grayscale values. From the test results, only FunFuseAn and MdAFuse have a more prominent hierarchical structure of grayscale values, which can retain MR feature information more completely, while the grayscale value hierarchical results of other methods are difficult to identify.

Finally, for the third green square area in Fig. 6, the amount of information is rich, including the choroid plexus, occipital horn of the lateral ventricle, occipital lobe, insular lobe and insula. Five arrows point to the result of our method, which is obviously different from those of the other methods. Area 1 is the choroid plexus. In the MR image, the whole location is two concentric circles with gray color. The grayscale value of the outer circle is deeper than that of the inner circle, and the difference is obvious. The grayscale value of the center of the circle is second to that of the outer circle, and it is also prominent. In the SPECT image, the grayscale value is more even. The results of U2Fusion and DeFusion show that the radius of the inner circle is relatively large, the center point of the circle is not obvious, and the pixel value of the inner circle area is not different. FunFuseAn, FusionDN, CDDFuse and our MdAFuse method can better retain the MR information and can clearly find the obvious difference between the inner circle and the outer circle and the center of the circle. Similarly, the other arrows point to areas that work better in our approach. The fusion methods based on deep learning can be mainly distinguished from their operation at the feature level. Nine quality evaluation metrics (EN, MI,



(a) MR Image (b) SPECT Image (c) FunFuseAn (d) FusionDN (e) U2Fusion (f) DeFusion (g) CDDFuse (h) MdAFuse

Fig. 6. Performance Comparisons: The quality of our MdAFuse method is comparable to those of existing DL-based methods. From (c) to (h) are FunFuseAn [26], FusionDN [29], U2Fusion [31], DeFusion [32], CDDFuse [30] and MdAFuse.

TABLE II
COMPARISON WITH SOTA METHODS IN QUANTITATIVE RESULTS OF MEDICAL IMAGE FUSION OF PAIRS OF 74 MR-SPECT AND 42 PAIRS OF MR-PET RESPECTIVELY.

MR-SPECT	EN	MI	SD	VIF	SCD	SSIM	CC	MSE	rSFe
FunFuseAn [26]	4.2348	2.6155	55.3604	0.8947	1.0452	1.5794	**0.8917**	0.0535	-0.3340
FusionDN [29]	3.9001	**2.6912**	56.5726	0.5216	0.9051	0.4167	0.8756	0.0388	-0.2623
U2Fusion [31]	4.0581	2.4404	42.2308	0.3188	1.7674	0.3016	0.8832	0.0344	-0.5399
DeFusion [32]	3.7698	1.7543	49.9900	0.5393	0.7180	1.4480	0.8654	0.0225	-0.3775
CDDFuse [30]	3.9151	2.4925	**58.3698**	0.9666	1.3458	1.4717	0.8381	0.0375	**-0.0449**
MdAFuse	**4.4135**	2.5343	58.0256	**1.0780**	**1.1483**	**1.6555**	0.8838	**0.0112**	-0.2508
MR-PET	EN	MI	SD	VIF	SCD	SSIM	CC	MSE	rSFe
FunFuseAn [26]	4.5540	2.5407	57.5225	0.7334	1.1915	1.4668	**0.8122**	0.1123	-0.3633
FusionDN [29]	4.2210	**2.6952**	64.5598	0.4297	1.4188	0.4453	0.8024	0.0756	-0.2153
U2Fusion [31]	4.4952	2.0185	50.1283	0.3383	1.5538	0.2741	0.7760	0.0453	-0.5511
DeFusion [32]	4.1463	1.6765	63.5379	0.5199	1.4278	1.4335	0.7909	3.3017	-0.2963
CDDFuse [30]	4.2248	2.0258	**70.7307**	0.7056	**1.6863**	1.4905	0.7958	0.0700	**-0.0316**
MdAFuse	**4.7241**	2.3880	60.5385	**0.8432**	1.2910	**1.5466**	0.8094	**0.0250**	-0.3086

SD, VIF, SSIM, SCD, CC, MSE, rSFe) are used to evaluate the quality of the fusion results which shown in the Fig. 7. From this figure, we can observe that our method has an advantage in 6 metrics compared to other methods, especially in SSIM, SCD, VIF, and MSE, where the difference from the sub-optimal value is quite pronounced. For the rSFe metric, the CDDFuse method has the smallest value, but this relatively small value seems a bit unusual. This might be due to its less preservation of information in SPECT images, as shown in Fig. 6(g). In comparison, our method performs better on this metric. Although our performance in the SD and MI metrics is not outstanding, the gaps from the maximum and second-maximum values are not substantial.

Quantitative results are presented in Table II. The values in the table represent the average performance of each method in various metrics. In Table II, values displayed in bold represent the best, and those underlined indicate the sub-optimal. Overall, our MdAFuse method outperforms others, showing superior results in the majority of metrics, particularly on EN, VIF, SSIM, and MSE. CDDFuse is currently the state-of-the-art method, and when applied to MR-SPECT fusion, we achieved improvements of 11.14% and 18.38% in VIF and SSIM, respectively. Additionally, we reduced MSE by 2.63%. In the case of MR-PET fusion, our method resulted in increases of 13.76% and 5.61% in VIF and SSIM, while reducing MSE by 4.50%. The variance in the calculation of different metrics for each method is provided in the Appendix. What's more, we also compared results of MR-CT pairs fusion with SOTA methods in the Appendix.

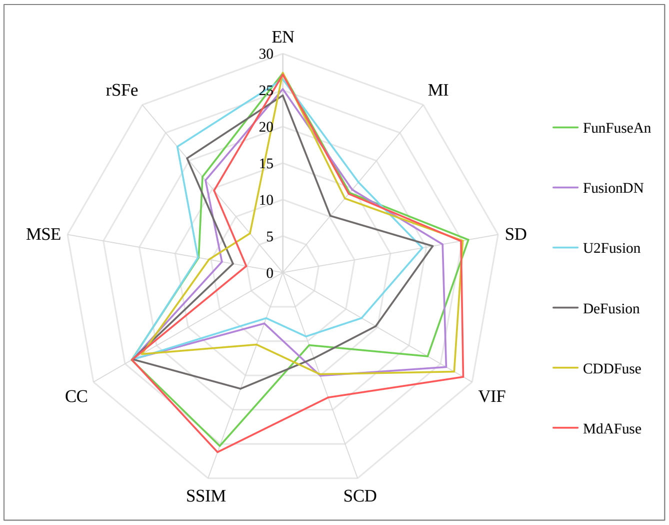


Fig. 7. Objective assessment of six kinds of deep-learning-based fusion methods on Fig. 6.

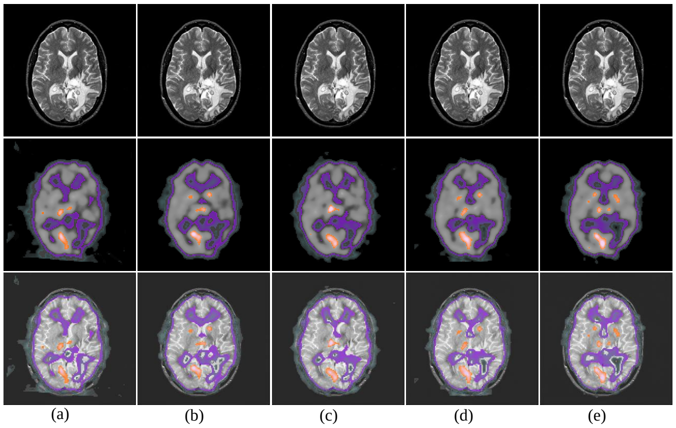


Fig. 8. Enhanced fusion effect of one group of MR-SPECT medical images sampled five time points from a certain case by time. ((a), (b), (c), (d) and (e) represent scans of subjects at five time points, including MR and SPECT images.)

C. Key feature enhancement test

The first two lines in Fig. 8 show MR-SPECT source images that were sampled in five periods from the same case with brain abnormalities. The white and black changes in the second line of SPECT images can be used to determine the location of the lesions and the severity of the disease, which is helpful for tumor location, identification and auxiliary diagnosis. The third line is the fusion result of the corresponding MR-SPECT image pairs of the first two lines, representing the coloring effect of MdAFuse. Each column from left to right represents the imaging results at a different time. In Fig. 8, we can see the change of energy information in SPECT and the key position to judge the abnormality easily, which is mainly reflected in the yellow contour and the purple area. From left to right, the number of yellow areas on fused result increases, the yellow area in the lower left parietal lobe increases, and the brightness of white areas also improves. If this brightness represents higher glucose energy, it is likely that diabetes and hyperglycemia may occur in this location.For the area with black spots, the number of black blocks decreases with time, but the area and concentration increase. For example, the occipital horn of lateral ventricle has only one small point in Fig. 8(a), and two small black blue areas are formed in Fig. 8(b). The area of the black blue area in Fig. 8(c) becomes larger and moves downward, while the two small areas are merged in Fig. 8(d), and finally the area is expanded in Fig. 8(e). It indicates that there is abnormality in this area. If the color region represents the cell metabolism ability, then the serious DeFusioniciency of glucose and protein may cause the corresponding diseases in the occipital horn of lateral ventricle. Therefore, coloring the fusion results can enhance the effect after fusion, which can reveal not only the abnormal tissue structure position but also the energy change, which is helpful for accurate diagnosis of the disease evolution.

TABLE III
RUNNING TIME OF IMAGE FUSION METHODS (UNIT: S).

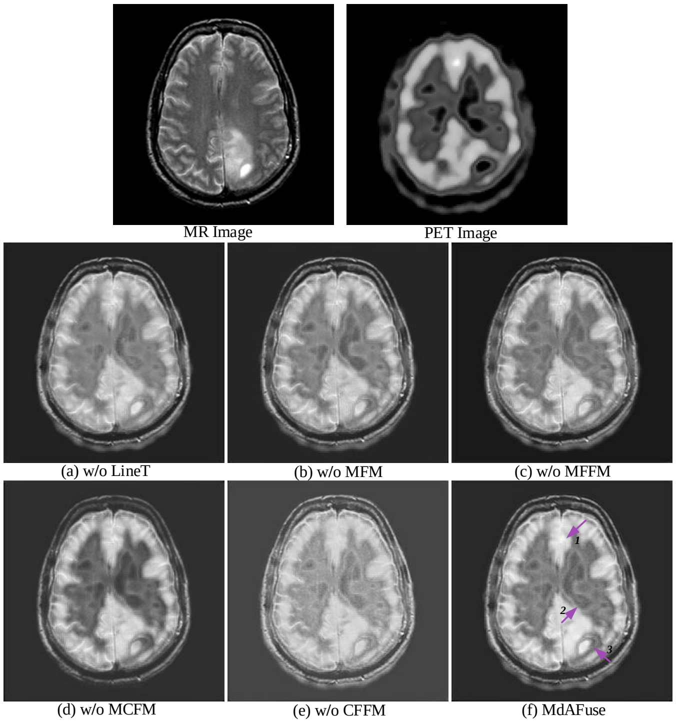
PCF	LatLrr	atsIF	DeFusion	U2Fusion
0.3495	12.7103	3.3208	0.3934	0.2157
FunFuseAn	FusionDN	CDDFuse	MdAFuse	
0.1706	0.3898	1.9163	1.2270	

D. Computational cost

Under our computation environment, 137 pairs of different data modalities are tested by nine kinds of methods. The average run-time for each method is recorded. All results are listed in Table III. We can see that the average run-time of our method (MdAFuse) is 1.2270 seconds, which is an acceptable time cost.

E. Ablation experiments

To fully illustrate the necessity and effectiveness of the fusion strategy for coarse feature extraction, fine feature extraction, multi-scale feature extraction and linear transformation, we conduct five groups of ablation experiments, the results are shown in Fig. 9. The first line in Fig. 9 represents a pair of multimodal (MR-PET) images, and the second to third lines show the fusion results without linear transformation operation (w/o LinearT), without the multi-scale feature extraction module (w/o MFM), without multi-scale and fine feature extraction module (w/o MFFM), without multi-scale and coarse feature extraction module (w/o MCFM), without coarse and fine feature extraction module (w/o CFFM), and the fusion strategy of our method proposed (MdAFuse). In Fig. 9(f), the three arrows contain different information in different results. In Area 1, the brightness information on PET should be retained. In Fig. 9(b) and Fig. 9(c), the brightness of this position is not obvious. Fig. 9(a) and In Fig. 9(d), the position is fuzzy, which may be caused by the low brightness intensity. Area 2 is a black area in the PET image, and the gray level is uneven in the MR image. In Fig. 9(b) and Fig. 9(d), it is an area with an even gray level. In Fig. 9(a) and Fig. 9(c), the gray level is arranged in a certain gradient, but the edge is smoothed out. In Fig. 9(f), the effect is better, and more



(a) w/o LineT (b) w/o MFM (c) w/o MFFM

(d) w/o MCFM (e) w/o CFFM (f) MdAFuse

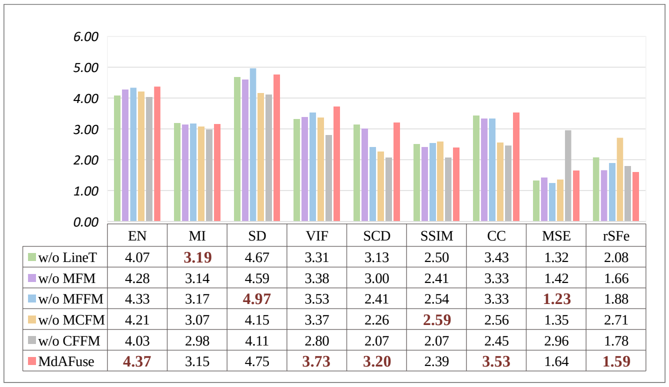
Fig. 9. Ablation experiment of loss one or more part of the proposed image fusion method.

structural information is retained. The area pointed by Arrow 3 on MR is a small area larger than a lentil with strong and even brightness, which has a hierarchical relationship with the surrounding color difference. In the six cases, MdAFuse can reflect the enhancement effect of this position, and good results are obtained.

Fig. 10 shows some values for quality assessment, which refer to the nine quality evaluation metrics corresponding to the result calculations. In Fig. 10, the lower part is the original six metric values and the upper part is the bar chart display. In order to better observe and compare different metric values in the same chart, linear transformation is adopted to each metric value, with the transformation coefficients shown in Table I. In Fig. 10, we can observe that our method has the best values in 5 metrics and the sub-optimal value in 1 metric. Therefore, comparatively speaking, our proposed method demonstrates the most advantageous performance. Namely, MdAFuse in this paper is best for other metrics, which can also prove that the network constructed by adding multi-dimensional features and using linear transformation can obtain better fusion results.

V. Conclusion

In this paper, we proposed a novel DL-based fusion framework for multi-dimensional features that combines spatial features and channel features. At the same time, a deep separation convolution network is used to excavate richer and useful information, which can provide significant features and rich detail features for source images for subsequent image comprehension and application analysis. Three different feature extraction modules and an adaptive linear fusion mechanism based on the correlation of each dimension feature are used to preserve the spatial texture information of MR images and the physiological metabolism information of PET/SPECT images. In addition, we also proposed a key feature enhancement method that can enhance the visualization of fusion images in different periods for the same case, which is helpful for clinical applications such as tumor localization, segmentation and disease tracking. Different diseases are evaluated by different commonly used imaging examinations, and multimodal medical image fusion is diverse. Follow-up work will continue to study different types of medical fusion methods and apply them to AI medical diagnosis. Although our method has demonstrated certain advantages in MR-PET/SPECT image fusion, there are still some limitations. Our method only focuses on key features of MR and PET/SPECT images without considering the specificity of a certain disease. In addition, for the fusion of brain images, there was no added neuroscience analysis for areas of potential disease. It will be a meaningful research direction to combine neuroscience and AI methods more deeply in the further research.



	EN	MI	SD	VIF	SCD	SSIM	CC	MSE	rSFe
🟩 w/o LineT	4.07	**3.19**	4.67	3.31	3.13	2.50	3.43	1.32	2.08
🟪 w/o MFM	4.28	3.14	4.59	3.38	3.00	2.41	3.33	1.42	1.66
🟦 w/o MFFM	4.33	3.17	**4.97**	3.53	2.41	2.54	3.33	**1.23**	1.88
🟨 w/o MCFM	4.21	3.07	4.15	3.37	2.26	**2.59**	2.56	1.35	2.71
⬜ w/o CFFM	4.03	2.98	4.11	2.80	2.07	2.07	2.45	2.96	1.78
🟥 MdAFuse	**4.37**	3.15	4.75	**3.73**	**3.20**	2.39	**3.53**	1.64	**1.59**

Fig. 10. Quality evaluation of the ablation experiment responding to Fig. 9

.

References

[1] Y. Nakamoto, K. Tamai, T. Saga, T. Higashi, T. Hara, T. Suga, T. Koyama, and K. Togashi, "Clinical value of image fusion from MR and PET in patients with head and neck cancer," *Molecular Imaging and Biology*, vol. 11, no. 1, pp. 46–53, 2009.

[2] W.-D. Heiss, "The potential of PET/MR for brain imaging," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 36, no. 1, pp. 105–112, 2009.

[3] T. Zhou, M. Liu, H. Fu, J. Wang, J. Shen, L. Shao, and D. Shen, "Deep multi-modal latent representation learning for automated dementia diagnosis," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, vol. 11767, 2019, pp. 629–638.

[4] T. Zhou, K.-H. Thung, Y. Zhang, H. Fu, J. Shen, D. Shen, and L. Shao, "Inter-modality dependence induced data recovery for MCI conversion prediction," in *Medical Image Computing and Computer Assisted Intervention*, vol. 11767, 2019, pp. 186–195.

[5] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," *arXiv preprint arXiv:1804.08992*, pp. 1–6, 2018.

[6] A. N. Fedorova and M. G. Zeitlin, "Quantum multiresolution: tower of scales," in *International Conference on Mathematical Modeling in Physical Sciences*, vol. 490, no. 1, 2014, pp. 012 216:1–012 216:5.

[7] D. Gambhir and M. Manchanda, "Waveatom transform-based multi-modal medical image fusion," *Signal Image and Video Processing*, vol. 13, no. 2, pp. 321–329, 2019.

[8] J. Du, M. Fang, Y. Yu, and G. Lu, "An adaptive two-scale biomedical image fusion method with statistical comparisons," *Computer Methods and Programs in Biomedicine*, vol. 196, pp. 105 603:1–105 603:13, 2020.

[9] J. R. Benjamin and T. Jayasree, "Improved medical image fusion based on cascaded PCA and shift invariant wavelet transforms," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 2, pp. 229–240, 2018.

[10] Y. Yang, M. Yang, S. Huang, Y. Que, M. Ding, and J. Sun, "Multifocus image fusion based on extreme learning machine and human visual system," *IEEE Access*, vol. 5, pp. 6989–7000, 2017.

[11] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Information Fusion*, vol. 42, pp. 158–173, 2018.

[12] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Medical image fusion via convolutional sparsity based morphological component analysis," *IEEE Signal Processing Letters*, vol. 26, no. 3, pp. 485–489, 2019.

[13] K.-j. Xia, H.-s. Yin, and J.-q. Wang, "A novel improved deep convolutional neural network model for medical image fusion," *Cluster Computing-Journal of Netwoorks Software Tools and Applications*, vol. 22, no. 1, pp. 1515–1527, 2019.

[14] H. Li, X. He, Z. Yu, and J. Luo, "Noise-robust image fusion with low-rank sparse decomposition guided by external patch prior," *Information Sciences*, vol. 523, pp. 14–37, 2020.

[15] K. Zhan, Q. Li, J. Teng, M. Wang, and J. Shi, "Multifocus image fusion using phase congruency," *Journal of Electronic Imaging*, vol. 24, no. 3, pp. 033 014:1–033 014:13, 2015.

[16] J. Wen, S. Xuan, Y. Li, Q. Gao, and Q. Peng, "Image-segmentation algorithm based on wavelet and data-driven neutrosophic fuzzy clustering," *Imaging Science Journal*, vol. 67, no. 2, pp. 63–75, 2019.

[17] J. Zhong, B. Yang, Y. Li, F. Zhong, and Z. Chen, "Image fusion and super-resolution with convolutional neural network," in *Chinese Conference on Pattern Recognition*, vol. 663, 2016, pp. 78–88.

[18] B. Rajalingam, F. Al-Turjman, R. Santhoshkumar, and M. Rajesh, "Intelligent multimodal medical image fusion with deep guided filtering," *Multimedia Systems*, vol. 28, no. 4, pp. 1449–1463, 2022.

[19] A. P. James and B. V. Dasarathy, "Medical image fusion: A survey of the state of the art," *Information fusion*, vol. 19, no. SI, pp. 4–19, 2014.

[20] Y. Xiaolong, L. I. Demin, T. U. Ya, and S. Baoci, "Identification of subthreshold depression based on deep learning and multimodal medical image fusion," *Chinese Journal of Medical Imaging Technology*, vol. 36, no. 8, pp. 1158–1162, 2020.

[21] B. Rajalingam and R. Priya, "Multimodal medical image fusion based on deep learning neural network for clinical treatment analysis," *International Journal of ChemTech Research*, vol. 11, no. 6, pp. 160–176, 2018.

[22] F. Zhao and W. Zhao, "Learning specific and general realm feature representations for image fusion," *IEEE Transactions on Multimedia*, vol. 23, pp. 2745–2756, 2021.

[23] B. Xiao, B. Xu, X. Bi, and W. Li, "Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion," *IEEE Transactions on Image Processing*, vol. 30, pp. 163–175, 2021.

[24] I. Shopovska, L. Jovanov, and W. Philips, "Deep visible and thermal image fusion for enhanced pedestrian visibility," *Sensors*, vol. 19, no. 17, pp. 3727:1–3727:21, 2019.

[25] R. Hou, D. Zhou, R. Nie, D. Liu, L. Xiong, Y. Guo, and C. Yu, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 640–651, 2020.

[26] N. Kumar, N. Hoffmann, M. Oelschlägel, E. Koch, M. Kirsch, and S. Gumhold, "Structural similarity based anatomical and functional brain imaging fusion," in *Multimodal Brain Image Analysis and Mathematical Foundations of Computational Anatomy*, 2019, vol. 11846, pp. 121–129.

[27] H. Xu, F. Fan, H. Zhang, Z. Le, and J. Huang, "A deep model for multi-focus image fusion based on gradients and connected regions," *IEEE Access*, vol. 8, pp. 26 316–26 327, 2020.

[28] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Information Fusion*, vol. 54, pp. 99–118, 2020.

[29] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 484–12 491.

[30] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. Van Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5906–5916.

[31] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 1, pp. 502–518, 2022.

[32] P. Liang, J. Jiang, X. Liu, and J. Ma, "Fusion from decomposition: A self-supervised decomposition approach for image fusion," in *European Conference on Computer Vision*, vol. 13678, 2022, pp. 719–735.

[33] H. Xu, J. Ma, and X.-P. Zhang, "MEF-GAN: Multi-exposure image fusion via generative adversarial networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 7203–7216, 2020.

[34] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 1383–1396, 2021.

[35] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[37] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2261–2269.

[38] Z. Li and R. T. Rajan, "Geometry-aware distributed kalman filtering for affine formation control under observation losses," in *International Conference on Information Fusion*, 2023, pp. 1–7.

[39] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *IEEE International Conference on Computer Vision*, 2021, pp. 1–30.

[40] J. W. Roberts, J. van Aardt, and F. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, pp. 023 522:1–:023 522:28, 2008.

[41] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[42] Y.-J. Rao, "In-fibre bragg grating sensors," *Measurement Science and Technology*, vol. 8, no. 4, pp. 355–375, 1997.

[43] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Information Fusion*, vol. 14, no. 2, pp. 127–135, 2013.

[44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[45] Aslantas, V, Bendes, and Emre, "A new image quality metric for image fusion: The sum of the correlations of differences," *Aeu-international Journal of Electronics and Communications*, vol. 69, no. 12, pp. 160–166, 2015.

[46] S. S. Han, H. T. Li, and H. Y. Gu, "The study on image fusion for high spatial resolution remote sensing images," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 7, pp. 1159–1164, 2008.

[47] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Research*, vol. 30, no. 1, pp. 79–82, 2005.

[48] Y. Zheng, E. A. Essock, B. C. Hansen, and A. M. Haun, "A new metric based on extended spatial frequency and its application to DWT based fusion algorithms," *Information Fusion*, vol. 8, no. 2, pp. 177–192, 2007.

Jinyu Wen received the M.Sc. degree in engineering from the Guangxi University for Nationalities, Nanning, China, in 2019. She is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. Her current research interests include machine learning, deep learning, and medical image analysis.

Asad Khan received the B.Sc. degree in applied mathematics from the Government College University Faisalabad, Faisalabad, Pakistan, in 2010, the M.Sc. degree in mathematical modeling and scientific computing from the Air University, Islamabad, Pakistan, in 2012, and the Ph.D. degree in image and video processing from the University of Science and Technology of China, Hefei, China, in 2017. He is currently a Postdoctoral Fellow and a Teaching Instructor with the Guangzhou University, Guangzhou, China. His current research interests include image processing, computer vision, deep learning, computational photography, hyperspectral imaging, and wearable computing.

Amei Chen received the M.Sc. degree in clinical specialty and the M.D. in radiology both from the Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, in 2003 and 2006, respectively. She is currently a Senior Doctor of Radiology with the Second Affiliated Hospital of South China University of Technology, Guangzhou, China. Her current research interests include brain imaging, artificial intelligence research of depression, and Parkinson's disease.

Weilong Peng received the Ph.D. degree in computer application technology from the Tianjin University, Tianjin, China, in 2017. He is currently a Lecturer with the Metaverse Research Institute, School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. His current research interests include image processing, computer vision, and deep learning.

Meie Fang received the Ph.D. degree in applied mathematics from Zhejiang University, Hangzhou, China. She is currently a Full Professor with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. She worked in the Institute of Computer Graphics and Image, Hangzhou Dianzi University from June 2007 to June 2017, and was transferred to Guangzhou University in June 2017. She has served as a Postdoctoral Fellow in the State Key Lab of CAD & CG, Zhejiang University and the Postdoctoral Station of Computer Application Technology, Shanghai Jiao Tong University. She visited City University of Hong Kong and Purdue University of the United States for the purpose of academic exchange several times in recent years. Her current research interests include computer graphics and medical image analysis.

C. L. Philip Chen (Fellow, IEEE) received the Ph.D. degree in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1988. He is currently a Chair Professor and the Dean of the School of Computer Science and Engineering, South China University of Technology, Guangzhou, China. Being a Program Evaluator of the Accreditation Board of Engineering and Technology Education (ABET) in the U.S., for computer engineering, electrical engineering, and software engineering programs, he successfully architects the University of Macau's Engineering and Computer Science programs receiving accreditations from Washington/Seoul Accord through Hong Kong Institute of Engineers (HKIE), of which is considered as his utmost contribution in engineering/computer science education for Macau as the former Dean of the Faculty of Science and Technology. He is a Fellow of IEEE, AAAS, IAPR, CAA, and HKIE; a member of Academia Europaea (AE), European Academy of Sciences and Arts (EASA), and International Academy of Systems and Cybernetics Science (IASCYS). He received IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings. He is also a highly cited researcher by Clarivate Analytics in 2018 and 2019. His current research interests include systems, cybernetics, and computational intelligence. Dr. Chen was a recipient of the 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University (in 1988), after he graduated from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA in 1985. He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of the IEEE Transactions on Cybernetics (2020-2021) and the IEEE Transactions on Systems, Man, and Cybernetics: Systems (2014-2019), and currently, an Associate Editor of the IEEE Transactions on Fuzzy Systems. He was the Chair of TC 9.1 Economic and Business Systems of International Federation of Automatic Control from 2015 to 2017, and currently is a Vice President of Chinese Association of Automation (CAA).

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 200 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, computational art, and creative media.