

MsgFusion: Medical Semantic Guided Two-Branch Network for Multimodal Brain Image Fusion

Jinyu Wen , Feiwei Qin , Jiao Du , Meie Fang , Xinhua Wei, C. L. Philip Chen , *Fellow, IEEE*,
and Ping Li , *Member, IEEE*

Abstract—Multimodal image fusion plays an essential role in medical image analysis and application, where computed tomography (CT), magnetic resonance (MR), single-photon emission computed tomography (SPECT), and positron emission tomography (PET) are commonly-used modalities, especially for brain disease diagnoses. Most existing fusion methods do not consider the characteristics of medical images, and they adopt similar strategies and assessment standards to natural image fusion. While distinctive medical semantic information (MS-Info) is hidden in different modalities, the ultimate clinical assessment of the fusion results is ignored. Our MsgFusion first builds a relationship between the key MS-Info of the MR/CT/PET/SPECT images and image features to guide the CNN feature extractions using two branches and the design of the image fusion framework. For MR images, we combine the spatial domain feature and frequency domain feature (SF) to develop one branch. For PET/SPECT/CT images, we integrate the gray color space feature and adapt the HSV color space feature (GV) to develop another branch. A classification-based hierarchical fusion strategy is also proposed to reconstruct the fusion images to persist and enhance the salient MS-Info reflecting anatomical structure and functional metabolism. Fusion experiments are carried out on many pairs of MR-PET/SPECT and MR-CT images. According to seven classical

objective quality assessments and one new subjective clinical quality assessment from 30 clinical doctors, the fusion results of the proposed MsgFusion are superior to those of the existing representative methods.

Index Terms—Brain image, feature extraction, image fusion, two-branch network, medical semantic information.

I. INTRODUCTION

IMAGE fusion has been widely applied in computer vision, remote sensing, traffic safety and other fields. Since the 1990 s, image fusion technologies have been developed and applied in the medical field. However, there are many differences between natural and medical images such as the signal-to-noise ratio, resolution, relevant area size and image scene [1], [2], [3], [4]. It is often inefficient to directly apply traditional natural image networks to medical images, because it will lead to a performance degradation. Therefore, we think that it is necessary to deeply analyze the characteristics of medical images and design a dedicated network model. Medical image fusion makes it convenient for doctors to observe and estimate a case and analyze lesions to make a more accurate diagnosis. In general, CT/MR/SPECT/PET are often used to observe brain diseases by doctors. Different medical image modalities have unique MS-Info. The fusion results of different modalities should persist and enhance the pertinent information (i.e., MS-Info) in each modality that is significant for diagnosis. Therefore, we propose a medical semantic guided dual branch network. In the feature extraction stage, image features are guided according to the MS-Info of different modalities, and the network branch strategy conducive to the extraction of the corresponding image features is adopted to ensure that the MS-Info of each modality is maintained and enhanced in the fusion result. Image fusion technology integrates and enhances this information from two source images into one image so that doctors can observe, comprehend and diagnose diseases more conveniently and accurately. Therefore, the fusion technology of brain CT/MR/SPECT/PET images is of great significance to help doctors diagnose brain diseases.

In recent years, many medical image fusion methods have been proposed. These existing methods mainly include two categories, i.e., traditional artificial fusion methods [5], [6], [7], [8], [9], [10] and deep learning based fusion methods [11], [12], [13], [14], [15], [16], [17], [18], [19], [20]. The traditional method is driven by artificial cognition. Traditional feature extraction methods mainly rely on manual extraction, which requires professional domain knowledge and complex parameter adjustment

Manuscript received 20 July 2022; revised 14 January 2023 and 4 March 2023; accepted 12 April 2023. Date of publication 8 May 2023; date of current version 18 January 2024. This work was supported in part by the National Natural Science Foundation of China under Grants 62072126, 61972121, and 62176071, in part by the Fundamental Research Projects Jointly Funded by Guangzhou Council and Municipal Universities under Grant 202102010439, in part by the Postgraduate Innovation Project of Guangzhou University under Grant 2021GDJC-D16, in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020015, in part by the Guangzhou Science and Technology Project under Grant 202102021243, 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 Grants P0030419, P0042740, P0044520, P0043906, and P0035358. The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Ngai-Man (Man) Cheung. (Jinyu Wen, Feiwei Qin, and Jiao Du contributed equally to this work.) (Corresponding author: Meie Fang.)

Jinyu Wen, Jiao Du, and Meie Fang are with the Metaverse Institute, School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou 511400, China (e-mail: wjy1361120721@163.com; dujiao19880429@126.com; fme@gzhu.edu.cn).

Feiwei Qin is with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: qinfeiwei@hdu.edu.cn).

Xinhua Wei is with the Department of Radiology, Second Affiliated Hospital of South China University of Technology, Guangzhou 510641, China (e-mail: eyxinhuawei@163.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, Kowloon, Hong Kong (e-mail: p.li@polyu.edu.hk).

Digital Object Identifier 10.1109/TMM.2023.3273924

processes. Moreover, each method is targeted at specific application scenarios. Deep learning based methods are data driven. They can obtain deeply abstract features by learning a large number of samples. The expression of the dataset is more efficient and accurate, and the extracted abstract features have strong robustness and generalization. Currently, deep learning methods have been successfully applied in many fields of image processing. However, deep learning based medical image analysis is still at an early stage of development. Corresponding studies are becoming hot topics and will have extensive application prospects. This article focuses on a new fusion method for CT/MR/SPECT/PET brain images, which aims to persist and enhance important MS-Info of the original images to assist doctors in diagnosing brain diseases. In sum, our work makes the following contributions:

- This is the first study to focus on the distinguished MS-Info of multimodal medical images and map them into corresponding image features. A medical semantic guided two-branch network is designed to effectively learn the deep features corresponding to the MS-Info of the different modalities (MR/PET/SPECT/CT).
- In the SF-branch, we propose a spatial-frequency combined feature extraction scheme to more conveniently extract the corresponding features of key MS-Info from MR images on the basis of preserving the original image information, which is the first time to do so in the neural network of medical image fusion.
- In the GV-branch, we propose a novel scheme to combine information from both the gray color space and the improved luminance component in the HSV color space to extract deep features corresponding to the key MS-Info of the PET/SPECT images on the basis of preserving the original image information.
- We propose a new clinical assessment that is derived from doctors. They assess fusion quality depending on how much MS-Info from source images was pertained and enhanced in their fusion image. Simultaneously, seven classical assessment indicators are adopted in this research. Many experiments illustrate that the proposed MsgFusion method is superior to nine kinds of representative fusion methods under all these assessment mechanisms.

The rest of the article is organized as: Section II introduces the related works. Section III describes the proposed image fusion method in detail. Section IV shows the experimental results and analysis. The conclusion is given in Section V.

II. RELATED WORK

Existing fusion methods include traditional and deep learning based methods. In this section, we mainly introduce related works of medical image fusion, the theory of Fourier transform and HSV color space transform in image processing.

A. Medical Image Fusion

Medical images have their own particularities, and medical image fusion plays an important role in image-guided medical diagnosis, treatment and other computer vision tasks. Among them, multimodal image fusion modules are also designed for

disease auxiliary diagnosis, such as [18], [21]. Algarni proposed a diagnosis system based on multimodal image fusion, which is suitable for fusing MR and CT images [22], however, it is only suitable for fusing MR and CT images. To develop a fusion method that can accurately preserve detailed information even if the image is damaged, Li et al. applied the low-rank sparse matrix dictionary learning method to medical image fusion [23], which can achieve the effect of denoising and enhancement at the same time. Panigrahy et al. proposed a medical image fusion method (WPADPCNN) by using dual channel PCNN [24], which can obtain a good fusion effect but is only suitable for MR-SPECT fusion. Parvathy proposes a fusion model based on optimized threshold and deep learning [25] that can provide anatomical and physiological data to experts to facilitate the diagnostic process. Focusing on pseudocolor images in the color space domain, a dual-scale image fusion method based on the Otsu adaptive threshold (atsIF) is proposed [8].

Hermessi et al. [26] proposed a multimodal MR and CT image fusion method based on similarity learning of a convolutional neural network. Das et al. proposed a new image fusion based on low-rank texture prior decomposition and super-pixel segmentation [27], which combined three kinds of schemes, i.e., gray wolf optimization, optimized low rank texture prior, and a pixel-related Gaussian mixing model to improve the visual fidelity of the fused image effectively. Kumar et al. proposed a new CNN method that is specifically used for MR and PET image fusion [14], and the structural similarity index (SSIM) was used as the loss function in the training process. Liang et al. [28] proposed a multilayer cascade fusion network (MCFNet). The network supplements the spatial information lost in the two downsampling processes of the fused medical image. Yu et al. [16] proposed the network (IFCNN) and Li et al. [17] proposed the universal convergence network (NestFuse). Both of these methods used a convolution neural network to extract features and then adopted the maximum value of features strategy for multimodal medical image fusion. They can obtain a good fusion effect, however, the loss of global texture information cannot be avoided because convolution networks cannot capture long-distance dependencies.

Currently, most medical image fusion methods have limitations. First, MR/CT/PET/SPECT are commonly used brain imaging technologies, however, only two modalities are considered in most medical image fusion methods. Second, most of the methods mainly follow the idea of natural image fusion. In fact, there are differences between medical images and natural images, and there are also large differences among different modalities of medical images. The key MS-Info of the different modalities is not fully considered. Third, the fusion strategy of the existing methods adopts the weighted summation strategy and the maximum strategy. The selection of the weight coefficient needs to be tested many times and determined by the result. The summation strategy will result in fuzzy fusion images, and the maximum strategy will lose some key features. Such a simple and single fusion strategy will lead to the loss of key MS-Info in the source images. Fourth, the ultimate purpose of medical image fusion is to facilitate the reading of scans by clinicians, and the quality of the fusion image will affect the doctors' diagnosis of diseases. Therefore, a quality assessment by

clinicians should be the gold standard. Existing medical image fusion methods take little account of clinician evaluations. Overall, the most important point is that the MS-Info of the source images is not fully considered. Therefore, this article constructs a novel fusion network focusing on MS-Info, the fusion target of brain medical images and clinician evaluations.

B. Application of Frequency Domain Transform in Image Processing

In the process of image processing, the image is often converted to frequency domain processing, commonly used Fourier transform. In the field of image processing, Fourier transform can be used to obtain the frequency distribution of the spatial images, and then various image processing in the frequency domain can purposefully achieve many functions. In recent years, Fourier transform in image coding [29], image detection [30], image compression [31], image analysis [32], [33], image registration [34] and image reconstruction [35] have wide applications. Fourier transform should have potential advantages in image fusion. Naidu et al. [36] proposed a fast Fourier transform (MFFT)-based algorithm for the pixel-level fusion of multiresolution images. However, this method is based on the fusion of multiresolution natural images. To the best of our knowledge, there is no method for medical image fusion using Fourier transform. In addition, frequency-domain analysis methods also have undeniable potential in deep learning-based methods. For example, Kai et al. [37] proposed a frequency-based learning method, which proved the universality and superiority of frequency-domain learning methods in the classification, detection and segmentation tasks. The processing of the image in the frequency domain saves image information effectively and improves the accuracy. Therefore, we also adopt frequency domain processing to process medical images and combine deep learning methods to achieve a better fusion effect.

Fourier transform is a very important algorithm in the field of digital image processing that can transform images from the spatial domain to the frequency domain. Essentially, the image is the same as the original image, and the amplitude and phase components, that is, the global and local information of the image, can be more intuitively analyzed. This is very important information for the global texture and local geometry shape in MR images, and represents the semantic features of the soft tissue edges and the internal structures. Processing MR image by Fourier transform plays a key role in the localization of lesions resulting from fusion.

C. Application of Color Space Transform in Image Processing

The RGB, YUV and HSV color spaces are commonly used in image processing. The RGB color space is the most commonly used image color representation space. YUV is easy to compress, facilitating transmission and processing. The HSV color space is most suitable for the visual perception of human beings because color changes in this space are easily distinguished by human beings. H (Hue) represents the hues, expressing the color preference of images, and S (Saturation) represents the intensity or purity of colors. V (Value) indicates the brightness

of a color. RGB and HSV conversion is often used in image processing to enhance the image [38]. (R, G, B) are the red, green, and blue coordinates of a color, respectively, whose values are real numbers between 0 and 1. If M is equal to the maximum of R, G and B , and m is equal to the minimum of these values, the conversion process between RGB and HSV is shown in (1):

$$\begin{aligned} H &= \begin{cases} 0^\circ & \text{if } M = m \\ 60^\circ \times \frac{G-B}{M-m} + 0^\circ, & \text{if } M = R \text{ and } G \geq B \\ 60^\circ \times \frac{G-B}{M-m} + 360^\circ, & \text{if } M = R \text{ and } G < B \\ 60^\circ \times \frac{B-R}{M-m} + 120^\circ, & \text{if } M = G \\ 60^\circ \times \frac{R-G}{M-m} + 240^\circ, & \text{if } M = B \end{cases} \\ S &= \begin{cases} 0, & \text{if } M = 0 \\ \frac{M-m}{M} = 1 - \frac{m}{M}, & \text{otherwise} \end{cases} \\ V &= M. \end{aligned} \quad (1)$$




HSV is more consistent with human visual characteristics than the RGB color space [39]. The use of multiple channels in the HSV color space can be handled separately and independently of each other. Therefore, the workload of image analysis and processing can be greatly simplified in the HSV color space. It is also used in image fusion [40]. The HSV color space is widely used in image processing, and different channels are usually used to address different problems. Compared with the RGB space, the HSV space can express the lightness and shade of color, tone and vividness very intuitively, which is convenient for the contrast between colors and the communication of feelings. Therefore, in this article, we also apply the advantages of the HSV color space to medical image fusion. To better extract useful MS-Info, we adopt a self-defined V component, which will be introduced in detail in the next chapter.

III. METHOD

Different modalities of medical images contain distinctive MS-Info, which is very important for clinical disease diagnosis. Therefore, we should persist and enhance the respective MS-Info of multimodal medical images in their fusion result. To achieve this goal, we first analyze the MS-Info of each modality according to clinical medicine and imaging theory. Then, we map the key MS-Info into the image features and design effective extraction strategies for the different features. The details of how the key medical semantic information of MR/CT/PET/SPECT guides the design of the two branches of MsgFusion are shown in Table I, which guides us to construct the two branches of the proposed MsgFusion.

The overall framework of MsgFusion is illustrated in Fig. 1, where extracting features after preprocessing and fusing are critical procedures. First, in the stage of deep MS-Info extraction, the network combines two feature extraction branches, the SF-branch and GV-branch, as shown in Fig. 1. Fourier transform used in the SF-branch and the HSV color space considered in the GV-branch not only makes full use of the spatial and frequency relationship but also extracts rich semantic features from the image's important information. Second, is our hierarchical

TABLE I
HOW KEY MEDICAL SEMANTIC INFORMATION OF MR/CT/PET/SPECT GUIDES THE DESIGN OF TWO BRANCHES OF MSGFUSION

Modality	Key MS-Info	Image Feature	Extraction Strategy	Branch
MR 	Clear shape of soft tissue	Boundary zone	High frequency band in Frequency Domain	SF-branch
	Clear internal structure of soft tissue	Internal texture detail	Low frequency band in Frequency Domain	
	NIL	More source image information	Spatial Domain	
CT 	High-resolution global anatomical structure of hard tissue	Global contour lines; Local shape location of tiny area; More source image information	Multi-scale; Concatenate; Gray color space	GV-branch
	High-resolution local anatomical structure of hard tissue			
PET 	Obvious display of early small lesion			
	High-distinguished functional-metabolic abnormality tissue	Brightness		

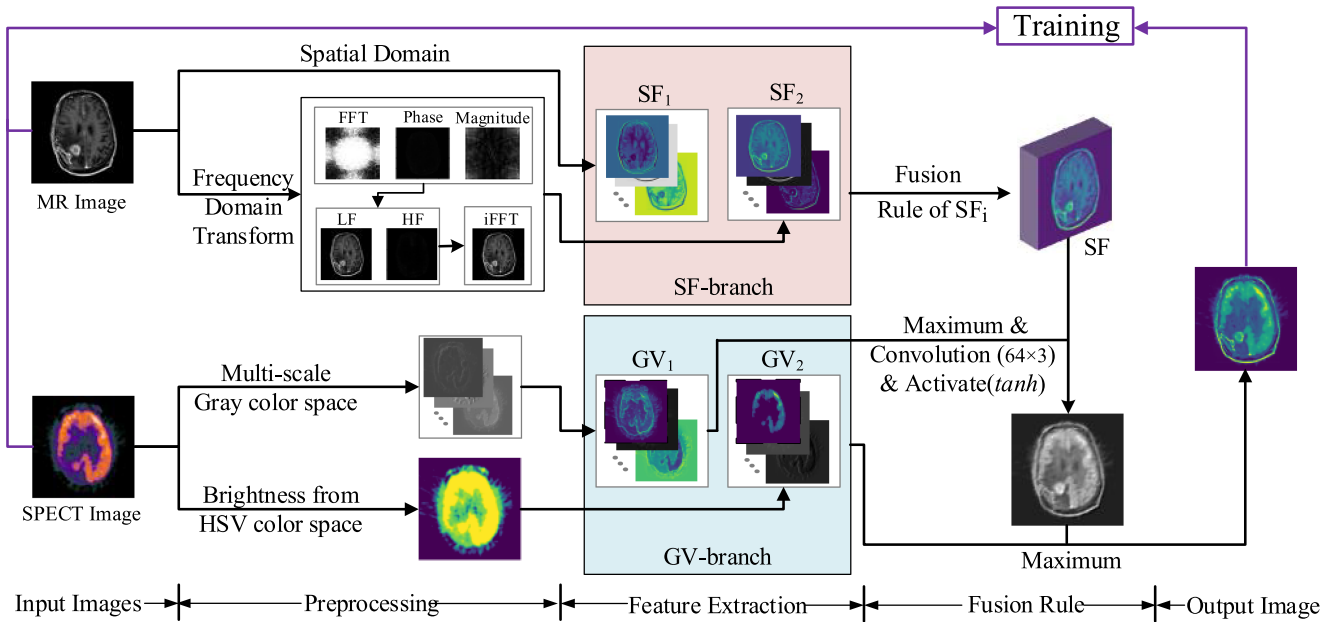


Fig. 1. The framework of our MsgFusion method: Medical Semantic Guided Two-Branched Network for Multimodal Brain Image Fusion. Two feature extraction branches proposed: SF-branch (the branch combines spatial domain and frequency domain to more conveniently extract the corresponding features of key MS-Info) and GV-branch (the branch combines gray color space and self-defined brightness information from HSV color space to enhance the extraction of corresponding features of key MS-Info). The fusion of SF-branch and GV-branch, we adopted the classification level fusion strategy: the anatomical structure features (SF and GV_1) were fused at the first level, and then the fusion results were fused with the functional metabolism features (GV_2) at the second level, which was beneficial to preserve and enhance key MS-Info of source medical images.

fusion strategy. The combination of these two feature extraction branches not only improves the performance of the algorithm but also efficiently obtains the important deep MS-Info of the multimodal medical brain images.

A. SF-Branch

The SF-branch is designed for extracting deep features from MR images. As illustrated in Table I, the MS-Info of MR, i.e., the clear shapes and internal structures of soft tissue, is more easily distinguished in the frequency domain as high- and

low-frequency band information. To effectively extract deep features corresponding to the MS-Info and more source image information, we adopt the strategy of combining the frequency domain with the spatial domain. Fig. 2 shows the procedure of the SF-branch.

The SF-branch of feature extraction includes two parts, one part is to obtain the deep features of the source MR image from the convolution characteristics of the neural network; the other part is to obtain the features corresponding to the MS-Info from the frequency domain. In the first part, the channel was amplified to 64, the size of the convolution kernel was set as

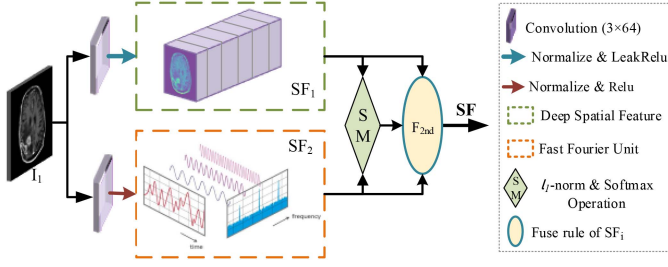


Fig. 2. Procedure of extracting SF-branch feature, which combines information from both the spatial domain and frequency domain to keep and enhance the key MS-Info of MR as much as possible.

7×7 , the stride size was set as 1 and the padding was set as 3. Batch normalization occurs when the convolution feature is obtained and activated using LeakyReLU ($\alpha = 0.2$, $\text{inplace} = \text{true}$). The output of this part is recorded as SF_1 , as shown in Fig. 2. In the other part, frequency domain processing was adopted first, and its output is recorded as SF_2 . An example of frequency domain processing is displayed in the square pointed to by an arrow marked with Frequency Domain Transform in Fig. 1. Concretely, the two-dimensional discrete Fourier transform and inverse transform of a piece of an $M \times N$ image are represented as (2) and (3). In the formula, x and y are image variables in the spatial domain, $f(x, y)$ represents the gray value at the point (x, y) , u and v are frequency variables, and when u and v are 0, it is the Fourier transform at the origin, which is equivalent to the average gray value of an image.

$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-j2\pi(ux/M + vy/N)}, \quad (2)$$

$$f(x, y) = \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u, v) e^{j2\pi(ux/M + vy/N)}. \quad (3)$$

If Re and Im are used to represent the real and imaginary parts of F , respectively, their calculation is according to (4) and (5) as:

$$\sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \cos\left(2\pi\left(\frac{wx}{N} + \frac{vy}{N}\right)\right), \quad (4)$$

$$\sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x, y) \sin\left(2\pi\left(\frac{wx}{N} + \frac{vy}{N}\right)\right). \quad (5)$$

After the image is transformed into the frequency domain by Fourier transform, every pixel is a complex number containing real and imaginary parts. The amplitude and phase of the image can be obtained by calculating the amplitude and phase of the complex number of each pixel. Then, the Fourier spectrum, the phase angle, and the amplitude are defined as:

$$P(u, v) = |F(u, v)|^2 = Re(u, v)^2 + Im(u, v)^2, \quad (6)$$

$$\phi(u, v) = \arctan\left[\frac{Im(u, v)}{Re(u, v)}\right], \quad (7)$$

$$|F(u, v)| = [Re(u, v)^2 + Im(u, v)^2]^{\frac{1}{2}}. \quad (8)$$

This is the first time medical image fusion has been performed using Fourier transform. The meaning of Fourier transform is to transform the gray distribution function of the image into the frequency distribution function of the image, and the inverse Fourier transform is to transform the frequency distribution function of the image into the gray distribution function. The amplitude and phase of the MR image can be obtained after Fourier transform. The amplitude of the image contains the global information of the image, namely, the texture information, while the phase contains the local information of the image, namely, the shape. Given the properties of Fourier transform, it can and does achieve a good fusion effect, which is illustrated by our proposed method.

When we are finishing the above steps, we need to fuse features between the spatial domain and frequency domain next. In this article, the weighted graph features are the fusion of the multiscale depth features to obtain the detailed structure of the spatial features. The weight mapping is done by the l_1 -norm and softmax operation via the objective function as:

$$\xi_k(x, y) = \frac{\|\psi_k(x, y)\|_1}{\sum_{i=1}^2 \|\psi_i(x, y)\|_1}, \quad (9)$$

where, $\|\cdot\|_1$ is the l_1 -norm, $k \in 1, 2$. (x, y) shows the corresponding position in the feature map, and each position denotes a dim dimensional vector in the deep features. ψ denotes a vector that has dim dimensions. The final fusion feature map, φ , is the superposition of two enhanced feature maps, which is represented by (10) as:

$$\varphi = \sum_{i=1}^2 \xi_k^{(i)}(x, y) \times \psi_i. \quad (10)$$

B. GV-Branch

The GV-branch is designed for extracting deep features from CT/PET/SPECT. On the one hand, by adopting a multiscale cascade strategy from the gray space, the GV-branch aims to extract global contour and local shape features from the source image and compensate for the loss of information at different scales. On the other hand, according to the analysis in Table I, the key MS-Info of PET/SPECT (highly distinguished functional metabolic abnormality tissue) is mainly reflected in the level of brightness of the image. To obtain the brightness information, we use an HSV color space transformation to compute the V component and improve it into a new luminance value to highlight the MS-Info of the functional images (PET/SPECT). The MS-Info of CT (high-resolution global and local anatomical structure of hard issue) and the part MS-Info of PET/SPECT (obvious display of early small lesion) are guided to multiscale strategies, concatenated in the gray color space. To capture information from different scales and layers with less information loss, multiscale and skip connection strategies are adopted. CT does not contain functional metabolic information. When CT is input into the GV-branch, only the deep features in the gray color space are needed. In total, to effectively extract the deep features corresponding to the MS-Info and more source image information, we adopt the strategy of combining the HSV color space with the gray color space.

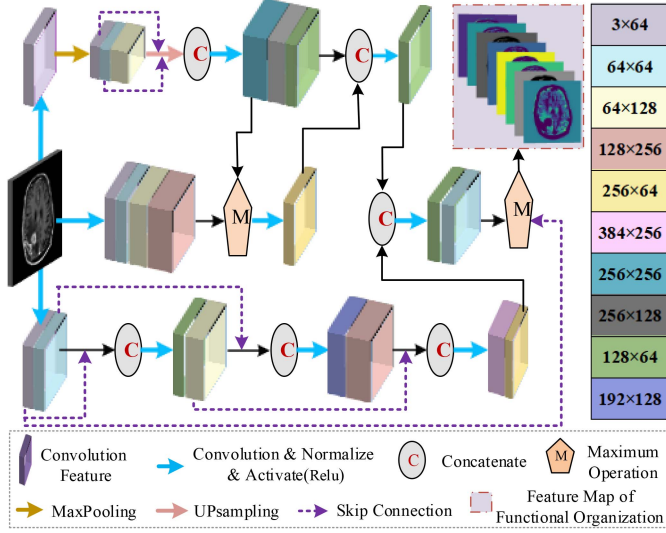


Fig. 3. First procedure of extracting feature of GV-branch feature, which adopts multiscale and concatenate strategies in the gray color space to extract both global contour and local shape features from source images.

1) *Obtaining Deep Convolution Features by a Multiscale Strategy:* Since convolution operations at different levels and resolutions can extract information with different importance degrees, a method of functional tissue feature extraction is proposed for brain images. The functional organization's feature branch network in this method is shown in Fig. 3. The network block mainly includes convolutional layer, pooling layer, up-sampling and skip connection operation. In Fig. 3, the specific steps to obtain the MS-Info are given. Different color cubes represent different numbers of input and output channels, which correspond to different thickness displays.

Multiple uses of concatenate can expand high-dimensional features and find more important semantic features more accurately. When multiple routes are merged, the maximum value method is used to highlight the high-frequency texture information. Because all kinds of brain diseases are uneven, using this feature extraction strategy can make the classification between the lesion degree and normal tissue more obvious. Then, we can more easily determine whether the local area is abnormal in the PET image, and observe clear contour and global structure information in the CT image. Moreover, the long and short skips are used to enhance the transfer of features with the green dotted line in Fig. 3, which can fully fuse different levels of visual features and reduce the loss of features in the process of feature transfer. This idea is inspired by DenseNet [41], which is mainly implemented by numerous dense network blocks. To solve the vanishing gradient problem, the network only cascades the weight coefficients of the front and rear layers of the convolutional operation, and does not cascade the pooling operation. In our network, we input the features of the previous layer and the previous multilayer to the next layer, and the connection modality is expressed as:

$$f_d = H_d([f_0, f_1, \dots, f_{d-1}]), \quad (11)$$

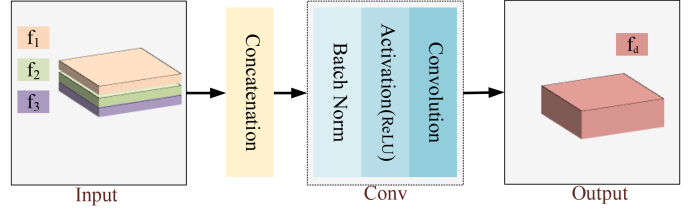


Fig. 4. Procedure of feature skip, which in order to capture long-term and multi-layer dependencies across regions and reduce the loss in the process of feature transfer.

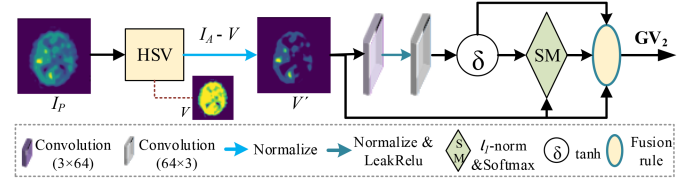


Fig. 5. Second procedure of extracting feature of GV-branch, which combines self-defined brightness information from the HSV color space to keep and enhance the key MS-Info of PET/SPECT as much as possible.

where, f represents the output, d represents the number of layers in the network and $H_d(*)$ represents the combination of nonlinear transfer functions, including the concatenate, BN, activation, and convolution operations. The parallel input of the multilayer features (not necessarily the adjacent layers) is operated by the combination function to obtain a new feature. The channel number of the new feature is determined by different convolution parameters. The detailed steps are shown in Fig. 4. This operation helps to capture long-term and multilayer dependencies across regions and to reduce the loss in the process of feature transfer. Therefore, a more complete structure of the brain image can be preserved.

2) *Obtain Brightness Information Via a Color Space Transform:* HSV (hue, saturation, brightness value) is a color space created according to the intuitive characteristics of colors, and is also known as the hexagonal cone model. The HSV color space can be processed separately and are independent of each other. Therefore, the workload of image analysis and processing can be greatly simplified in the HSV color space. Moreover, it is more consistent with human visual characteristics than the RGB color space. In this feature extraction part, we calculate the V component of the HSV space of the functional image to extract its luminance features for subsequent processing, as shown in Fig. 5. In this figure, I_p represents the original RGB image, I_A represents the RGB image, V represents the brightness component after transforming I_p into the HSV space. To further enhance the local information of the areas with obvious brightness, we define a new luminance value V' , which is shown in Algorithm 1.

As illustrated in Fig. 5, I_p is a piece of the PET image. Three distinct color blocks (yellow, light green and dark blue) can be found in it, which indicate areas with high energy or vigorous metabolism. Lesions most likely exist in such areas. These color blocks can be discerned from the foreground of I_p mainly because of their brightness information. However, from the

Algorithm 1: New Luminance Computation

Input: RGBA image $I_p : (R, G, B, A)$
Output: New luminance value V'

- 1: **newLum** (I_p):
- 2: $I_A \leftarrow \text{Compress}(I_p)$;
- 3: $I_{HSV} : (H, S, V) \leftarrow \text{Conver}(I_p)$;
- 4: $V' \leftarrow I_A - V$;
- 5: **return** V' ;
- 6: **Compress** (I_p):
- 7: $I_A \leftarrow I_p : (R) + I_p : (G) + I_p : (B)$;
- 8: **return** I_A ;
- 9: **Conver** (I_p):
- 10: $I_A \leftarrow \text{Compress}(I_p)$;
- 11: $I_{HSV} \xleftarrow{\text{Eq. (1)}} I_A$;
- 12: **return** $I_{HSV} : (H, S, V)$;

brightness image V , we can see that these meaningful areas are not yet clear enough for accurately locating the lesions. In Algorithm 1, we compute the difference image through $I_A - V$ to further highlight these color blocks. Then, the lesions can be clearly located on the right side of the V' image.

Based on the new luminance image, the deep features GV_2 in the HSV color space are extracted through the subsequent procedures, as illustrated in Fig. 5.

C. Classification-Based and Hierarchical Fusion and Reconstruction

To make full use of the deep MS-Info features extracted from the two branches, we adopted a classification-based and hierarchical fusion strategy, as shown in the fusion rule phrase of Fig. 1. The anatomical structure features (SF and GV_1) were fused at the first level, and then the fusion results were fused with the functional metabolism features (GV_2) at the second level, which was beneficial to preserve and enhance the key MS-Info of the medical images to be fused. Our fusion strategy is not a simple concatenate or a weighted sum. To maintain and enhance the significant texture characteristics of the multimodal images, we use the maximization method. At the same time, we use the convolution and activation functions to obtain the local detail structure features. As shown in Fig. 1, \tanh is adopted as the activation function.

D. Loss Function and Training

Due to the confidentiality of medical data, a large amount of multimodal medical data is not easy to obtain. Usually, many network models trained based on ImageNet data, such as ResNet [42] and VGG [43], are chosen to produce the pre-training parameters. In a quest for a better performance in solving the shattering gradient problem and convergence rate, we choose ResNet101 to migrate the last convolution feature on ImageNet, that is, to use the first layer of ResNet101 pre-trained on the ImageNet data as our first convolution layer to extract the valid image features. Since the pre-training model is designed for the classification task, the parameters of the first layer will

be adjusted according to our task in our network structure to achieve a better fusion effect. To obtain a more accurate reconstruction of the source image, we combine the structural and pixel information to construct a loss function as:

$$L = \omega L_S + (1 - \omega) L_P. \quad (12)$$

The total loss function consists of structural similarity L_S and pixel loss L_P , and ω is an adjustment coefficient whose value is determined by the training and testing effects according to different values. I_f represents the result of a fusion, $I_i, i = 1, 2$ represents the input image and the source image. $SSIM$ represents the structural similarity of two images [44]. For the two types of images, the structural similarity is calculated and combined. Assuming that the structural similarity of the two images is $F_{SSIM_i}, F_{SSIM_i} = 1 - SSIM(I_f, I_i), i = 1, 2$, and the total structural similarity function is calculated as follow:

$$L_S = \sum_{i=1}^2 \alpha_i \left(1 - \frac{(2\mu_{I_f}\mu_{I_i} + c_1)(2\sigma_{I_f}\sigma_{I_i} + c_2)}{(\mu_{I_f}^2 + \mu_{I_i}^2 + c_1)(\sigma_{I_f}^2 + \sigma_{I_i}^2 + c_2)} \right), \quad (13)$$

where, μ_{I_f} is the mean of I_f and μ_{I_1} and μ_{I_2} are mean values of I_1 and I_2 respectively. $\mu_{I_f}^2$ is the variance of I_f , and $\mu_{I_1}^2$ and $\mu_{I_2}^2$ are the variances of I_1 and I_2 respectively. $\sigma_{I_f I_1}$ and $\sigma_{I_f I_2}$ are the covariance between the original image and fused image, and c_i are constants. The pixel loss is the l_2 -norm. Assuming that the pixel losses of the two images are $F_{M_i}, F_{M_i} = \|I_f - I_i\|_2$, the total pixel loss function is calculated as follow:

$$L_P = \beta \sum_{j=1}^n (I_{f_j} - I_{1_j})^2 + (1 - \beta) \sum_{j=1}^n (I_{f_j} - I_{2_j})^2. \quad (14)$$

In (13) and (14), $\alpha_i, i = 1, 2, \beta$ are weight parameters for users to set in the interval $[0, 1]$. In this article, the three parameters are all set to 0.5. In (12), the value of $\omega \in [0, 1]$ is determined by the training and testing effects according to different values. When ω is taken as 0.7, the best result is obtained. Therefore, we choose this value to test the fusion effect in the experiment. In the training process, the initial learning rate is 0.001, 250 epochs, the batch size is 2, we employ batch normalization, and the optimization function is Adam. In addition, we adopted an adaptive loss adjustment strategy to update the learning rate to 0.1 times its original value every 50 epochs because a smaller learning rate is used to adjust the weight to obtain a better weight in the training process. Our approach has been implemented in Python 3.7.9 under the PyTorch version 1.5.0 with its corresponding CUDA version 10.1.

IV. EXPERIMENTAL RESULTS

To further prove the effectiveness of our proposed fusion method, experiments were carried out on different pairs of brain images (MR-SPECT/MR-CT/MR-PET). A total of 555 MR-PET image pairs as the training set were obtained from ADNI (<http://adni.loni.usc.edu/data-samples/access-data/>), and their sizes were 256×256 . Thirty pairs were used in the test MR-CT, MR-SPECT and MR-PET image pairs from the Whole Brain Atlas (<http://www.med.harvard.edu/aanlib/>), which are

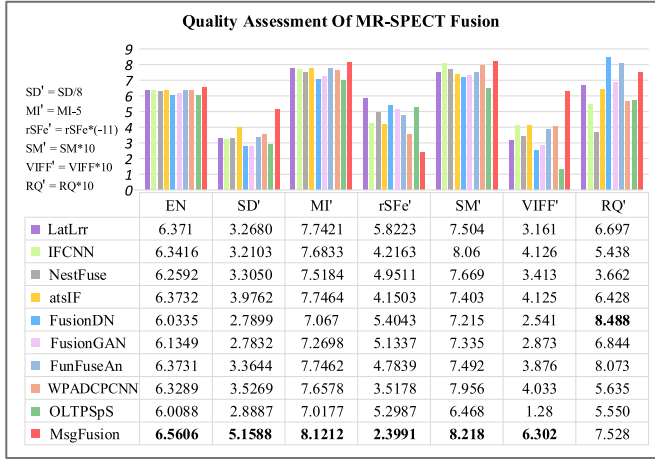


Fig. 6. Histogram of evaluation index values of different methods on MR-SPECT fusion. EN [45], SD [46], MI [47], SM [48], VIFF [49], RQ [50] means entropy, standard deviation, mutual information, structural similarity, visual information fidelity of fusion and edge information respectively. The larger the value is, the better the fusion effect is. rSFe [51] reflects of improved spatial frequency. The smaller its absolute value is, the better the fusion effect is. Except RQ, other six evaluation indexes reflect that MsgFusion achieves best fusion effect.

cerebrovascular disease (stroke) and tumor disease (brain tumor). All experiments were carried out on the GNU/Linux x86_64 system of the GeForce RTX 3090 Ti 12 Intel Core Interl(R) Xeon(R) CPU E5-2678 v3 2.50 GHz 64 GB RAM device. We have released the source code on GitHub through the link <https://github.com/22385wjy/MsgFusion>.

For each pair of experimental images, we use the proposed MsgFusion and nine other representative kinds of methods, i.e., LatLrr [7], IFCNN [16], NestFuse [17], atsIF [8], FusionDN [19], FusionGAN [15], FunFuseAn [14], WPADPCNN [24] and OLTPSpS [27] to produce the fusion results. We adopt six kinds of often-used assessment indices, including EN [45], SD [46], MI [47], rSFe [51], SM [48], VIFF [49] and a relatively novel index $R_Q^{(F/(AB))}$ (RQ for short) [50], to evaluate the fusion effect of these ten kinds of methods. RQ reflects how much edge information is perserved in the fusion image by computing fractional order differentiation, which substitutes the noise-sensitive Sobel operator used in [52]. Three sigmoidal functions (*tanh*, *arctan*, *logistic*) are adopted to obtain three metrics in that article. Considering their common monotonicity of function, we only choose one metric based on the *logistic* function to apply in our tests. Due to the large difference between the values of the different indicators, we made suitable linear transformations that are marked on the top left of Figs. 6 and 7 for the convenience of comparing them in the same figure.

A. Fusion of MR-SPECT Pairs

The first column in Fig. 8 shows the source MR and SPECT images. The MR images clearly show the texture details of the cerebrospinal fluid and other soft tissues. The different colors and brightness in SPECT can reflect the metabolic information. Fig. 8(a)–(i) show the results of the ten considered fusion methods. Fig. 8(j) shows the fusion results obtained by MsgFusion.

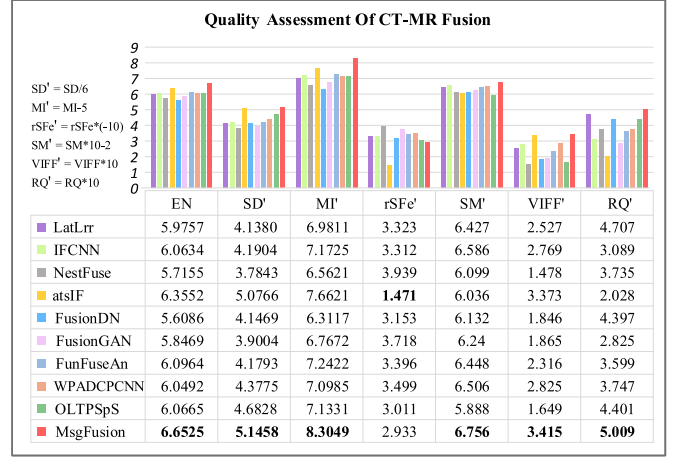


Fig. 7. Histogram of seven evaluation indexes of ten kinds of fusion methods corresponding to MR-CT fusion images displayed in Fig. 9. Our MsgFusion occupies six best index values, only its rSFe is sub-optimal.

Focusing on the regions indicated by the seven arrows, one can see that clearer structure and texture details are displayed than the corresponding regions in fusion images obtained by the other seven methods. Combining clinical information, we analyzed whether the MS-Info of MR and SPECT was well retained in the fusion results for the 7 regions indicated by the arrows in Fig. 8(j) inside the blue rectangle frame.

There is a nodule of the cerebellum in Region 1. A small white triangle with a clear contour can be seen in the MR source image, with a slight blurring of the upper right corner. However, in the SPECT source image, there is no obvious feature. In the fusion result of LatLrr, the image appears vague and indistinguishable from the surrounding tissues. In the fusion results of IFCNN, NestFuse and FusionDN, they all retain much more information from MR than SPECT. The enhancement effect of the upper right corner is not obvious. The small triangle in the results of FusionGAN and FunFuseAn is distorted. In the fusion result of OLTPSpS, the brightness information is significantly lost.

Region 2 is the left cerabellar hemisphere, and the features mainly come from the features in SPECT, which appear as a bright spot with a fuzzy edge, and it looks like a small black hole in the MR image. In the fusion results of LatLrr, IFCNN, NestFuse atsIF, WPADPCNN and OLTPSpS, we cannot easily find the bright spot. In the fusion results of FusionGAN and FunFuseAn, the bright spot is enhanced, but it is slightly blurry. In the fusion result of FusionDN, this bright spot is obvious, but it is integrated with the surrounding organizations, and the specific location cannot be determined. In the fusion result of MsgFusion, there is obvious brightness and contours, and the position is accurate.

Regions 3 and 6 represent the right cerabellar hemisphere. In the MR image, it is a region with an uneven distribution of gray values, which is shaped like a leaf. In the SPECT image, there are three adjacent gray hollow rings. From the SPECT image, we can find that there is a clear gray hollow ring at the position indicated by Arrow 6. It can be recognized in the fusion images obtained by NestFuse and atsIF, but is not very clear.

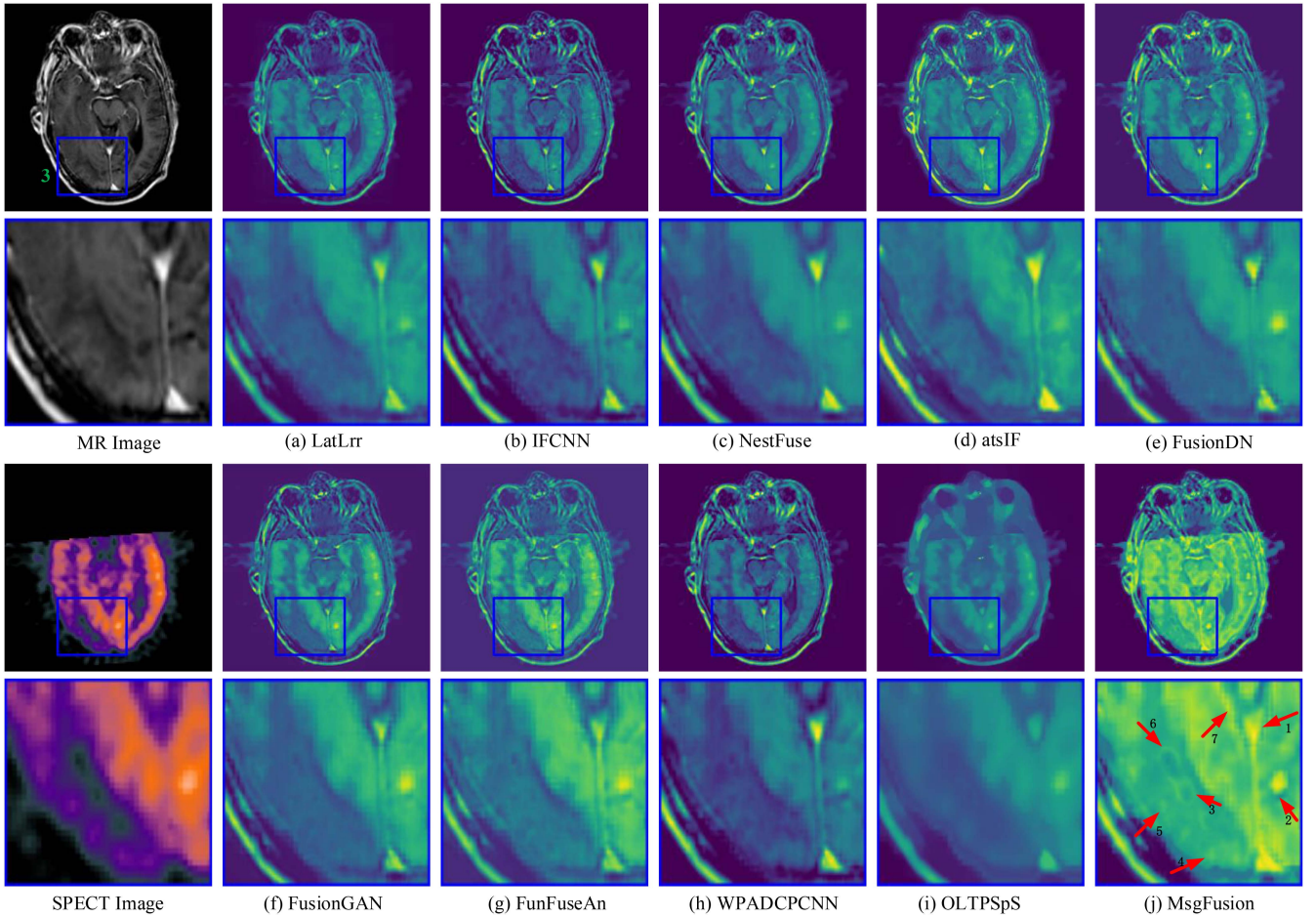


Fig. 8. Performance comparison of different methods (LatLrr [7], IFCNN [16], NestFuse [17], atsIF [8], FusionDN [19], FusionGAN [15], FunFuseAn [14], WPADPCNN [24], OLTPSpS [27], and our MsgFusion) for brain MR-SPECT image fusion. Focusing on regions pointed by seven arrows in the locally enlarged image(h), one can see that clearer structure and texture details are displayed than the corresponding regions in fusion images obtained by other seven methods. Both the MS-Info of MR and SPECT are well retained and enhanced. More detailed medical interpret refers to the corresponding text.

The ring at Region 3 disappears. In the fusion results of IFCNN, FusionDN, FusionGAN, FunFuseAn and WPADPCNN their rings at Region 6 are clearer, but not at Region 3. In our approach, several of the rings can be shown well, and it is easy to find their position.

Regions 4 and 5 are the margins of the occipital bone and right cerabellar hemisphere, respectively. This region in the MR appears lacy, but is not easily detected. It appears as a small point of inconspicuous brightness in SPECT. Comparatively speaking, MsgFusion preserves both the edges of MR and can easily find several points in SPECT.

Region 7 is the edge of the right cerabellar hemisphere near the nodules of the cerebellum. In the MR image, which is represented by light white edge information, there seems to be one straight line in the SPECT image. For the different results of all the methods, only MsgFusion can clearly find the edge.

Fig. 6 shows the assessment indices of the fusion results in Fig. 8. Different color curves refer to different index values, and each node represents different fusion methods. The last column is our approach, and we can find the advantages of our approach in the different indices. The EN, SD, MI, rSFe and VIFF of our approach are the best, so our approach has the most advantages.

Hence, our fusion results have a more suitable brightness, clearer contours and finer texture. Further more, our results persist and enhance important medical information. When abnormalities appear in the brain, we can determine the possible types of diseases by observing the color and brightness information in SPECT, and combining the texture of MR for an effective analysis.

B. Fusion of MR-CT Pairs

The first column in Fig. 9 shows the original CT and MR images. It is easy to find hard contours such as bone on CT images, and soft tissue structures on MR images. On the right are the results of the different fusion methods. Fig. 9(j) shows the fusion result and locally enlarged image obtained by MsgFusion. Let us focus on the six ROIs indicated by the arrows.

Region 1 represents the inferior horn of the lateral ventricle, which is dark gray on the CT images but is bright and clear on the MR images. The area of Region 1 in the MR images is large, so it has important MS-Info and is often used to judge whether there lesions exist. In the fusion results of LatLrr, IFCNN, FusionDN, FunFuseAn and OLTPSpS, the contour is not obvious and is not easy to find. In the atsIF result, the brightness of the position

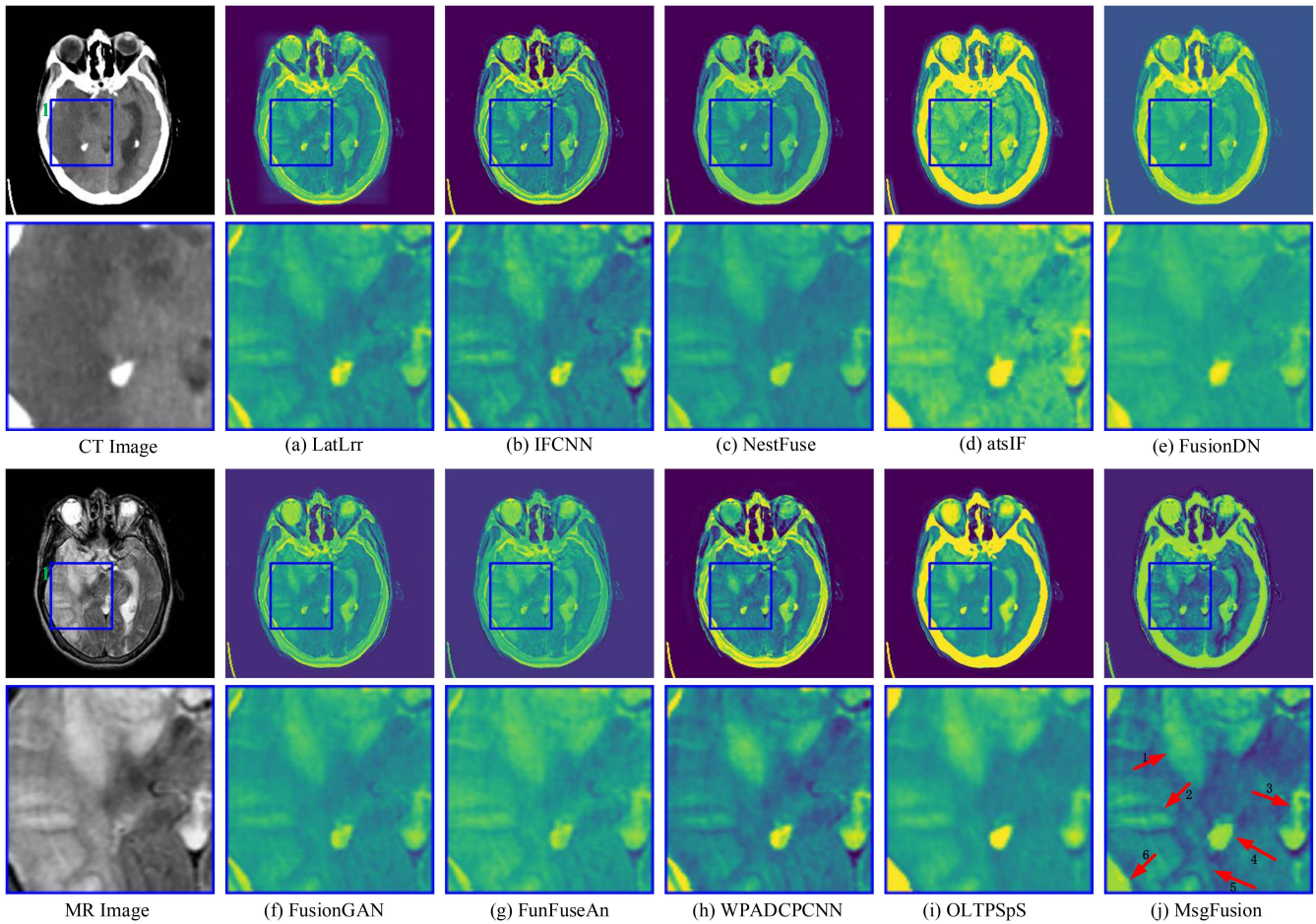


Fig. 9. Performance comparison of different methods (LatLrr [7], IFCNN [16], NestFuse [17], atsIF [8], FusionDN [19], FusionGAN [15], FunFuseAn [14], WPADPCNN [24], OLTPSPs [27] and our MsgFusion) for brain MR-CT image fusion. From fusion results and ROIs pointed by six arrows, one can see that MsgFusion shows best fusion effect. Both high-resolution global and local anatomical structure of hard tissue, and clear shape and internal structure of soft tissue are better retained and enhanced in its fusion result than others. More detailed medical interpret refers to the corresponding text.

is improved, but the edge is smoothed. In the fusion results of FusionGAN and WPADPCNN, the edge classification of the position is obvious, but the brightness is not sufficient. In the fusion result of this article, we can easily find the region, and its boundary.

Region 2 is insular, which is mainly reflected on the MR images. The area at this region is not as full as shown in the normal brain image, which indicates tissue loss of water or atrophy. In Fig. 9, we can find that only the region in the fusion result of MsgFusion is obvious and displays a clear contour distinguished from the other tissues.

Region 3 is the fourth ventricle in the CT image, and part of the information of the cerebro pontile and basilar artery is also connected on the MR image. The brightness of the atsIF method has good image contrast of bright and dark, however, the structural completeness of this region is destroyed. The results of the others in Region 3 are clear but not bright enough, while WPADPCNN and MsgFusion are better.

Region 4 is the cornu posterius ventriculi lateralis, which is mainly reflected on the CT images. It is not particularly obvious on the MR images, however, it can be discovered through a careful differentiation from the surrounding tissues. For LatLrr,

IFCNN, NestFuse and FusionGAN, the outline of Region 4 can be found. However, artifacts exist near its boundary, and the brightness information is insufficient. For FusionDN, this position is obvious, but the surrounding tissues are lost. The FunFuseAn fusion results show that the region becomes fuzzy. The result of OLTPSPs has enough brightness but the area is slightly smaller. Region 4 in our approach not only has a clear contour and an obvious boundary but also has sufficient brightness. The important medical features of the cornu posterius ventriculi lateralis in our approach persisted and were enhanced.

Region 5 is the central sulcus, which is mainly shown on the MR images. The density and brightness information of this location can help doctors determine whether abnormalities exist. For the MsgFusion region, which has a clear profile and brightness information, it is easy to find.

Region 6 is the parietal bone, which is mainly reflected in the CT image. When there are edge defects or incompleteness in this region, we can judge whether there is a brain injury. From the above analysis, we can determine that MsgFusion has a relatively better fusion effect. Furthermore, Fig. 7 shows the quality assessment values corresponding to Fig. 9. It shows that all the index values for MsgFusion except rSfE are the best.

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

LatLrr	IFCNN	NestFuse	atsIF	FusionDN
41.2933	0.2241	0.9762	6.7982	0.6873
FusionGAN	FunFuseAn	WPADCPCNN	OLTPSpS	MsgFusion
0.4924	0.3599	39.6705	48.59	2.0048

This means that the information of the original CT and MR not only persists well but is also enhanced in the fusion result of our approach. When intracerebral hemorrhage can be determined by observing the MR density and area, the general area of the cerebral hemorrhage can be found in CT images.

C. Computational Cost

Under our computation environment, 30 pairs of different data modalities are tested by ten kinds of methods. The average run-time for each method is recorded. All results are listed in Table II. We can see that the average run-time of our method is 2.0048 seconds. Although it is not a fast method, the time cost is acceptable.

D. Questionnaire Survey

For medical image fusion, the ultimate goal is to provide doctors with easily observed fusion images for qualitative, quantitative and locational analysis of lesions. Therefore, this subjective evaluation can confirm the advantages of the MSG method in a clinical sense. To prove the clinical effectiveness of our algorithm, we conducted an online questionnaire survey on the fusion effect, which we distributed to thirty doctors from the Neurology Department and Medical Imaging Department of different hospitals. These doctors' clinical experiences were more than 10 years (15 doctors), between 5 and 10 years (3 doctors), between 3 and 5 years (6 doctors) and less than 3 years (6 doctors). In this questionnaire, we designed 15 questions based on 15 groups of fusion experiments. For each question, there were 6 pieces of fusion images produced by 6 kinds of representative methods (i.e., LatLrr, NestFuse, atsIF, FusionGAN, FunFuseAn and MsgFusion) as options. The order of the fusion images from the different methods was randomly arranged. Each respondent was permitted to select one or two options with the best fusion results as answers for each question. Doctors did not need to identify how much the fused images retain the original image information. They only needed to judge which fusion results are more conducive to their observation and clinical diagnosis based on their clinical experience. In the end, we received valid answers from 29 participants.

The statistical results are shown in Table III, in which the times selected for the fusion images from each method are recorded. In each column marked with $Q_i (i = 1, \dots, 15)$, the following numbers correspond to the times selected by doctors for the fusion images of the six methods. In 15 groups of experiments, the fusion images produced by MsgFusion are most frequently selected as the best fusion images in 8 groups, and second frequently selected as the best fusion images in 4 groups. From the view of clinical doctors, the fusion effect of MsgFusion far

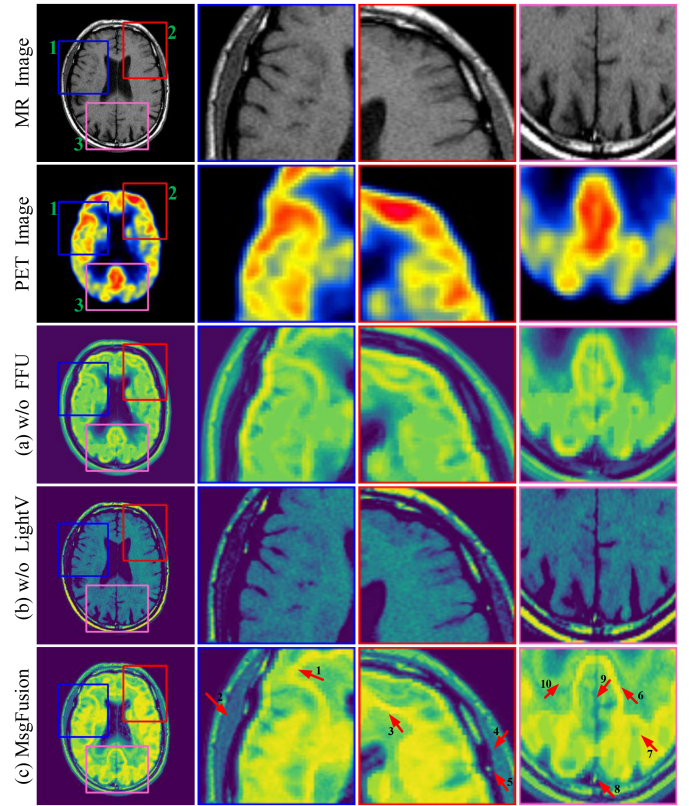


Fig. 10. Results of the ablation experiments. (a) Without frequency domain processing. (b) Without improved brightness from the HSV color space. (c) Our proposed MsgFusion.

exceeds that of any other considered method. The last column marked with \sum lists the total times for each kind of fusion method whose fusion images were selected. The calculation shows that the times selected by our method are 26.5% and 10.17% higher than those of the suboptimal method.

E. Ablation Study

To illustrate the necessity and effectiveness of combining the frequency domain in the SF-branch and HSV color space in the GV-branch, we performed ablation experiments on the MR-PET fusion. The experimental results are shown in Fig. 10. The first two rows show the source images of the MR and PET and their respective three locally enlarged ROIs. The following three rows show the fusion results obtained by MsgFusion without frequency domain processing, MsgFusion without considering improved brightness from HSV color space, and the proposed MsgFusion. In the last row of locally enlarged images, we mark ten arrows to point to regions with obvious medical characteristic information. Arrows 1 and 3 point to the frontal lobe, Region 2 is the frontal bone, Region 4 is the facies interna ossis frontalis, Region 5 is the space between the frontal bone and frontal lobe, the region of the parietal lobe is indicated by Arrows 6, 7 and 10, Region 8 is the superior sagittal sinus, and Arrow 9 points to the mediastinum cerebri. When FFT is not involved in the fusion process, the fusion results can preserve the structural features in PET more completely but will lose more MR image

TABLE III
TIMES SELECTED OF FUSION IMAGES FROM EACH METHOD IN THE QUESTIONNAIRE SURVEY

	Q ₁	Q ₂	Q ₃	Q ₄	Q ₅	Q ₆	Q ₇	Q ₈	Q ₉	Q ₁₀	Q ₁₁	Q ₁₂	Q ₁₃	Q ₁₄	Q ₁₅	Σ
LatLrr [7]	6	12	9	5	3	3	7	4	8	4	7	4	10	6	10	98
NestFuse [17]	16	10	7	9	8	10	8	6	7	1	6	5	5	6	2	106
atsIF [8]	9	7	7	5	10	2	9	8	12	5	9	15	3	10	4	115
FusionGAN [15]	7	8	7	0	6	4	2	10	4	15	3	2	11	7	2	88
FunFuseAn [14]	0	1	1	6	4	4	5	4	9	7	4	6	6	1	12	70
Our MsgFusion	7	5	12	17	12	21	13	12	4	11	14	10	9	13	11	172

TABLE IV
SIX EVALUATION INDICES OF THREE KINDS OF FUSION METHODS (W/O FFU, W/O LIGHTV, AND MSGFUSION) CORRESPONDING TO MR-PET FUSION IMAGE DISPLAYED IN FIG. 10

Whole image	EN	SD	MI	rSFe	SM	VIFF
w/o FFU	4.5027	56.8141	9.0054	-0.5018	0.2704	0.3219
w/o LightV	3.5768	36.7745	7.1537	-0.2977	0.3523	0.1262
MsgFusion	4.476	67.5634	8.952	-0.4103	0.3205	0.4076
Average ROIs	EN	SD	MI	rSFe	SM	VIFF
w/o FFU	6.9201	52.0514	13.8401	-0.4204	0.5594	0.3227
w/o LightV	5.9916	34.0278	11.9832	-0.3802	0.6118	0.1738
MsgFusion	6.9598	54.9381	13.9195	-0.3467	0.5830	0.3652

information. When only FFT is used without considering the improved V' component of the HSV color space, the fusion results can preserve relatively complete MR image features but cannot reflect the functional information of PET. The fusion results distinctly become much better when both are considered, as shown in the fifth row of Fig. 10.

Table IV shows the corresponding calculation results of the six evaluation indices in Fig. 10. The red value in the table is the optimal value, and the blue value is the next best value. The evaluation indices of the entire fusion image were calculated, and it was found that, the method in this article had a better effect (the optimal one had two indices, and the second-optimal one had four indices). In addition, we also calculated the average values of the evaluation indices of the three ROI regions shown in Fig. 10, as shown in the next four rows of Table IV. From the average value of the evaluation indices of the three ROI regions, it can also be found that the method in this article is better than the other two methods (without Fourier and without a V component calculation).

V. CONCLUSION

In this article, a deep feature fusion approach for brain disease images guided by MS-Info, MsgFusion, is proposed. We analyze the key MS-Info of MR/CT/PET/SPECT to obtain its corresponding image features, and then find the most efficient extraction strategies. Therefore, a two-branch network is designed, including the SF-branch and GV-branch. The SF-branch combines the spatial domain and frequency domain information and the GV-branch combines the multiscale gray images and brightness from the HSV color space. The dual network mechanism successfully improves the generalization ability of CNN and fully reflects the importance of the frequency domain information and color space information to ensure the effectiveness of the fusion results. The medical brain images are processed and analyzed, including the MR-CT image fusion, MR-SPECT image fusion, and MR-PET image fusion. Experiments show that,

compared with the existing methods, our approach has obvious advantages. We also asked clinical doctors to evaluate the fusion results via a questionnaire survey. The statistical data also proved that the proposed MsgFusion achieves the best fusion effect. In the future, we will consider extending the framework to integrate CT, MR, PET, SPECT, DTI and two or more other imaging modalities and apply them to clinical diagnosis.

REFERENCES

- [1] T. Wei et al., "Beyond fine-tuning: Classifying high resolution mammograms using function-preserving transformations," *Med. Image Anal.*, vol. 82, 2022, Art. no. 102618.
- [2] Y. Shen et al., "An interpretable classifier for high-resolution breast cancer screening images utilizing weakly supervised localization," *Med. Image Anal.*, vol. 68, 2021, Art. no. 101908.
- [3] J. Guo, Z. Zhou, and L. Wang, "Single image highlight removal with a sparse and low-rank reflection model," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 282–298.
- [4] W. Xia, E. C. S. Chen, S. E. Pautler, and T. M. Peters, "A global optimization method for specular highlight removal from a single image," *IEEE Access*, vol. 7, pp. 125976–125990, 2019.
- [5] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Inf. Fusion*, vol. 35, pp. 81–101, 2017.
- [6] H. Li and X.-J. Wu, "Multi-focus noisy image fusion using low-rank representation," 2018, *arXiv:1804.09325*.
- [7] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," 2018, *arXiv:1804.08992*.
- [8] J. Du, M. Fang, Y. Yu, and G. Lu, "An adaptive two-scale biomedical image fusion method with statistical comparisons," *Comput. Methods Programs Biomed.*, vol. 196, 2020, Art. no. 105603.
- [9] M. Diwakar, P. Singh, and A. Shankar, "Multi-modal medical image fusion framework using co-occurrence filter and local extrema in NSST domain," *Biomed. Signal Process. Control*, vol. 68, 2021, Art. no. 102788.
- [10] B. Wang et al., "Latent representation learning model for multi-band images fusion via low-rank and sparse embedding," *IEEE Trans. Multimedia*, vol. 23, pp. 3137–3152, 2021.
- [11] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
- [12] Y. Liu et al., "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, 2018.
- [13] X. Yan, S. Z. Gilani, H. Qin, and A. Mian, "Unsupervised deep multi-focus image fusion," 2018, *arXiv:1806.07272*.
- [14] N. Kumar et al., "Structural similarity based anatomical and functional brain imaging fusion," in *Proc. Multimodal Brain Image Anal. Math. Found. Comput. Anatomy*, 2019, pp. 121–129.
- [15] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, 2019.
- [16] Y. Zhang et al., "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, 2020.
- [17] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
- [18] T. Zhou et al., "Deep multi-modal latent representation learning for automated dementia diagnosis," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 629–638.

- [19] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 12484–12491.
- [20] F. Zhao and W. Zhao, "Learning specific and general realm feature representations for image fusion," *IEEE Trans. Multimedia*, vol. 23, pp. 2745–2756, 2021.
- [21] T. Zhou et al., "Inter-modality dependence induced data recovery for MCI conversion prediction," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2019, pp. 186–195.
- [22] A. D. Algarni, "Automated medical diagnosis system based on multi-modality image fusion and deep learning," *Wireless Pers. Commun.*, vol. 111, no. 2, pp. 1033–1058, 2020.
- [23] H. Li, X. He, D. Tao, Y. Tang, and R. Wang, "Joint medical image fusion, denoising and enhancement via discriminative low-rank sparse dictionaries learning," *Pattern Recognit.*, vol. 79, pp. 130–146, 2018.
- [24] C. Panigrahy, A. Seal, and N. K. Mahato, "MRI and SPECT image fusion using a weighted parameter adaptive dual channel PCNN," *IEEE Signal Process. Lett.*, vol. 27, pp. 690–694, 2020.
- [25] V. S. Parvathy, S. Pothiraj, and J. Sampson, "A novel approach in multi-modality medical image fusion using optimal shearlet and deep learning," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 4, pp. 847–859, 2020.
- [26] H. Hermessi, O. Murali, and E. Zagrouba, "Convolutional neural network-based multimodal image fusion via similarity learning in the shearlet domain," *Neural Comput. Appl.*, vol. 30, no. 7, pp. 2029–2045, 2018.
- [27] M. Das, D. Gupta, P. Radeva, and A. M. Bakde, "Optimized multimodal neurological image fusion based on low-rank texture prior decomposition and super-pixel segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, 2022, Art. no. 5010409.
- [28] X. Liang, P. Hu, L. Zhang, J. Sun, and G. Yin, "MCFNet: Multi-layer concatenation fusion network for medical images fusion," *IEEE Sensors J.*, vol. 19, no. 16, pp. 7107–7119, Aug. 2019.
- [29] K. Abiko, K. Uruma, M. Sugawara, S. Hangai, and T. Hamamoto, "Image segmentation based graph-cut approach to fast color image coding via graph Fourier transform," in *Proc. IEEE Vis. Commun. Image Process.*, 2019, pp. 1–4.
- [30] C.-C. Kuo et al., "Fast Fourier transform combined with phase leading compensator for respiratory motion compensation system," *Quantitative Imag. Med. Surg.*, vol. 10, no. 5, pp. 907–920, 2020.
- [31] A. Gnutti, F. Guerrini, R. Leonardi, and A. Ortega, "Symmetry-based graph fourier transforms: Are they optimal for image compression?," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1594–1598.
- [32] L. Fang et al., "Joint demosaicing and subpixel-based down-sampling for bayer images: A fast frequency-domain analysis approach," *IEEE Trans. Multimedia*, vol. 14, pp. 1359–1369, 2012.
- [33] G. Hu, B. Cui, and S. Yu, "Joint learning in the spatio-temporal and frequency domains for skeleton-based action recognition," *IEEE Trans. Multimedia*, vol. 22, pp. 2207–2220, 2020.
- [34] T. Atta-Fosu and W. Guo, "Joint segmentation and nonlinear registration using fast Fourier transform and total variation," in *Research in Shape Analysis*, A. Genctav et al., Eds. Cham, Switzerland: Springer, 2018, pp. 111–132.
- [35] E. Pezzotti, "Efficient non-uniform fast fourier transform (NuFFT) implementation for MRI processing on FPGA," Ph.D. dissertation, Univ. Illinois at Chicago, Chicago, IL, USA, 2017.
- [36] V. P. S. Naidu, "Multi-resolution image fusion by FFT," in *Proc. Int. Conf. Image Inf. Process.*, 2011, pp. 1–6.
- [37] K. Xu et al., "Learning in the frequency domain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1737–1746.
- [38] P. Li, Y. Huang, and K. Yao, "Multi-algorithm fusion of RGB and HSV color spaces for image enhancement," in *Proc. Chin. Control Conf.*, 2018, pp. 9584–9589.
- [39] X. Jin et al., "Multimodal sensor medical image fusion based on nonsub-sampled shearlet transform and S-PCNNs in HSV space," *Signal Process.*, vol. 153, pp. 379–395, 2018.
- [40] J. Liu and M. Wei, "Scene sparse recognition method via intra-class dictionary for visible and near-infrared HSV image fusion," *J. Comput. Appl.*, vol. 38, no. 12, pp. 3355–3359, 2018.
- [41] G. Huang, Z. Liu, L. v. d. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.
- [42] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, pp. 4278–4284.
- [43] L. Wang, S. Guo, W. Huang, and Y. Qiao, "Places205-VGGNet models for scene recognition," 2015, *arXiv:1508.01667*.
- [44] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [45] J. W. Roberts, J. v. Aardt, and F. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 023522.
- [46] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, pp. 355–375, 1997.
- [47] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
- [48] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. IEEE Int. Conf. Image Process.*, 2003, vol. 3, pp. 173–176.
- [49] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, 2013.
- [50] A. Sengupta, A. Seal, C. Panigrahy, O. Krejcar, and A. Yazidi, "Edge information based image fusion metrics using fractional order differentiation and sigmoidal functions," *IEEE Access*, vol. 8, pp. 88385–88398, 2020.
- [51] 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," *Inf. Fusion*, vol. 8, no. 2, pp. 177–192, 2007.
- [52] C. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.



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



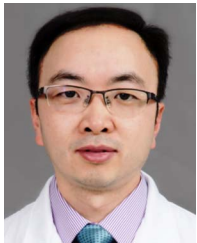
Feiwei Qin received the Ph.D. degree in computer science and technology from Zhejiang University, Hangzhou, China, in 2014. He is currently an Associate Professor with the School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou. His research interests include artificial intelligence, machine learning, image processing, and computer-aided design.



Jiao Du received the Ph.D. degree in computer science from the Chongqing University of Posts and Telecommunications, Chongqing, China, in 2017. She is currently a Lecturer with the School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, China. Her research interests include pattern recognition, deep learning, image processing, and computer vision.



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 was with the Institute of Computer Graphics and Image, Hangzhou Dianzi University, Hangzhou, from June 2007 to June 2017, and was transferred to Guangzhou University in June 2017. She was a Postdoctoral Fellow with the State Key Lab of CAD & CG, Zhejiang University and the Postdoctoral Station of Computer Application Technology, Shanghai Jiao Tong University, Shanghai, China. She visited City University of Hong Kong, Hong Kong, and Purdue University, West Lafayette, IN, USA, for the purpose of academic exchange several times in recent years. Her research interests include intelligent computer graphics, geometric deep learning, and medical image analysis.



Xinhua Wei received the M.Sc. degree in radiology from Guangxi Medical University, Nanning, China, in 2003, and the M.D. degree in radiology from Capital Medical University, Beijing, China, in 2007. He is currently a Professor of radiology with the Second Affiliated Hospital of South China University of Technology, Guangzhou, China. From May 2013 to June 2014, he was with the Department of Radiology, Wayne State University, Detroit, MI, USA, as a Senior Visiting Scholar. His research interests include brain imaging and artificial intelligence research of depression and Parkinson's disease.



C. L. Philip Chen (Fellow, IEEE) received the graduation degree from the University of Michigan at Ann Arbor, Ann Arbor, MI, USA in 1985, and 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) with 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), Hong Kong, 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 also a highly cited Researcher by Clarivate Analytics in 2018 and 2019. His research interests include systems, cybernetics, and computational intelligence. He is a Fellow of 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 was the recipient of the IEEE Norbert Wiener Award in 2018 for his contribution in systems and cybernetics, and machine learnings, 2016 Outstanding Electrical and Computer Engineers Award from his alma mater, Purdue University in 1988, after his graduation. He was the IEEE Systems, Man, and Cybernetics Society President from 2012 to 2013, the Editor-in-Chief of IEEE TRANSACTIONS ON CYBERNETICS during 2020–2021, and the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS during 2014–2019, and currently, an Associate Editor for 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.



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 authored or coauthored many 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 research interests include image/video stylization, colorization, artistic rendering and synthesis, realism in non-photorealistic rendering, computational art, and creative media.