ABSTRACT  Medical imaging plays a pivotal role in the clinical diagnosis of brain disease. There are many imaging methods to detect the state of tissues in the brain. While these imaging methods have advantages, they also have shortcomings. For example, magnetic resonance imaging (MRI) contains structural information but no functional characteristics of tissue, while positron emission tomography (PET) possesses functional characteristics but no structural information. The attention mechanism has been widely used in image fusion tasks, such as fusion of infrared and visible images and medical images. However, those attention models lack a balance mechanism for multimodal image features, affecting the final fusion performance. This paper proposes an end-to-end multimodal brain image fusion framework, MMI-fuse. Specifically, we first apply an autoencoder to extract the features of source images. Then, an information preservation weighted channel spatial attention model (ICS) is proposed to fuse the image features. We set an adaptive weight according to the information preservation degree of features. Finally, we use a decoder model to restructure the fused medical image. The proposed method increased the quality of fused images and decreased the fusion time effectively by the help of the improved attention model and encoder-decoder structure. To validate the performance of the proposed method, we collected 1590 pairs of multimodal brain images from the Harvard dataset and performed extensive experiments. Seven methods and five metrics were selected for the comparison experiments. The results demonstrate that the proposed method achieved notable performance on both the visual quality and objective metric score among these seven approaches. Moreover, the proposed method takes the least time among all compared methods.

INDEX TERMS  Medical image fusion, channel-spatial attention, MRI, PET, multimodal medical image.

I. INTRODUCTION
Image fusion aims to integrate information from different image sources into a single image. This technology is widely used in military, medicine, and civilian industries. In the clinical diagnosis domain, medical imaging plays a critical role in the diagnosis of cancer and tumors. On the other hand, the imaging mechanisms and expression formats of medical images are multidimensional. For example, computed tomography (CT) and magnetic resonance imaging (MRI) are both gray images with structural information: the former focuses on dense structures such as bones and implants; the latter focuses on soft tissues such as viscera and fat. Position emission tomography (PET) and signal-photon emission computed tomography (SPECT) are pseudocolor images with functional information: the former provides information on tissue based on disease status; the latter contains information on metabolism of organs, to be used for detection.

Switching between different modality images may reduce diagnosis efficiency and increase faults of doctors in deducing the clinical diagnosis results. With the development of image processing, fusion of multimodal medical images have become a promising and meaningful procedure that provides pivotal iconographic information.

In the past few decades, a handful of approaches have been proposed for medical image fusion. We briefly summarize these methods below. Methods based on multiscale decomposition (MSD) [1]–[3] follow a decomposition-reconstruction paradigm. They divide the source image into several layers, which express the information from different perspectives. These methods first utilize a transform function, such as pyramid transform [5], wavelet transform, contourlet
transform [2], or Shearlet transform [7], to decompose the input images into multiscale coefficients. Then, a manually designed fusion strategy is applied to fuse these coefficients. Finally, the fused coefficients are reconstructed by an inverse transform. The MSD methods can preserve the pseudocolors in PET and SPECT but limit the resolution of images. The performance of these methods mainly depends on the manually designed fusion strategy, which may fail to achieve good generation abilities. Methods based on subspace and component substitution (SCS) can be divided into two groups: statistical-based methods [8], [9] and color-based methods [11], [12]. The statistics-based approaches try to determine the hidden salient structures with higher order statistics, such as PCA and the Bayesian framework. These methods, however, are highly time-consuming and produce noisy results. The color-based methods first transform one source image into another color space. Then, a certain component is replaced or fused with another source image. Although they can preserve the structural information, the functional information of medical images is lost. The sparse representation (SR) came from the information, the functional information of medical images source image. Although they can preserve the structural information, the functional information of medical images is lost. The sparse representation (SR) came from the breakthrough of neuroscience that the human visual system is lost. SR-based methods have achieved promising results. However, SR-based methods have achieved promising performance in medical image fusion [14], [15], [16], [38]. Xiaosong Li et al. [14] proposed a fusion framework to solve the noise-free image fusion and noise- perturbed image fusion problem that adaptively designed the sparse reconstruction error parameter according to the noise level. This paradigm efficiently decreased the number of hyperparameters and strengthened the generalization ability of the model. However, SR-based methods still have drawbacks. The performance of SR-based methods highly depends on dictionary learning, and learning a complete dictionary is time-consuming. To improve the low computational efficiency of dictionary learning, Zhou et al. [4] proposed a method that reinforced the weak information of images by extracting and adding their multilayer details to generate informative patches.

In addition, fuzzy logic-based methods [17], morphology methods [18], and shallow machine learning methods [19] have also achieved remarkable performance. Although the methods mentioned above have made advances, the main shortage lines in feature extraction. Moreover, the multiscale decomposition and SR applied the handcraft feature extraction makes the poor generation ability. In recent years, deep learning (DL) has achieved great developments in natural language processing (NLP) and computer vision (CV) due to its excellent feature extraction and representation abilities. The medical image fusion methods based on DL, therefore, have better feature extraction ability than the traditional methods. Generative and adversarial networks (GAN) view image fusion as a game process. They apply a generator and a discriminator to create fake data and classify images as if they are real images to generate images. Ma et al. [20] first introduced the GAN into fusing infrared and visible images. Kang et al. [21] proposed a tissue-aware conditional generative adversarial network (TA-cGAN) that focused on the balance between color information in PET and anatomical information in MRI by joining spectral and structural losses. Additionally, the authors of TA-cGAN treated the image fusion as a Mix-Max optimization problem for the generator and discriminator. Although the performance of GAN was prominent, the training challenges, such as good initialization, still limited the fusion results. Recurrent neural networks (RNN) were mainly applied in speech recognition and text analysis. However, the local and global contextual dependency retrieval capabilities helped the RNN model learn the relation between each pixel representation and augment the representation of each pixel. Motivated by the local and nonlocal self-similarity of natural images, Zhao et al. [10] proposed a multi-focus image fusion model based on RNN. This model made use of the ability of the RNN model to retrieve long distant dependencies, which alleviated the information loss when the model went deeper. Convolutional neural networks (CNNs) made remarkable achievements in medical image analysis since they were proposed by LeCun et al. [41]. Many methods, [42], [43], have been proposed based on CNNs. The first fusion method based on CNN was proposed by Liu et al. [13] in 2017. They assumed image fusion to be a classification task. In reference [6], a fusion method based on a weighted parameter adaptive dual-channel PCNN (WPADCPCNN) was proposed. This method decomposed the source images into high-pass subbands and low-pass subbands. Then, the low-pass subbands were fused using a weighted multiscale morphological gradients-based rule and high-pass subbands with WPADCPCNN. However, the fusion performance of the dual-channel PCNN was limited by the parameters and the number of layers. Moreover, the high computational cost limited the scope of its application. Although RNN- and CNN-based methods have achieved advances, those fusion approaches require careful design of the efficient depth model. The performance of the fusion model is greatly affected by the depth of the model.

In this paper, we propose an unsupervised end-to-end framework, named MMI-Fuse, for multimodal medical brain image fusion. This framework effectively avoids the influence of neural network depth on the performance by using an autoencoder-decoder network. To preserve the information, we designed adaptive weights for the attention model, to be calculated according to the gradients of features. The proposed MMI-fuse includes an autoencoder-decoder model and a multi-attention fusion model. The main contributions regarding the proposed fusion network are summarized as follows:

1. We proposed a novel multimodal brain image fusion framework, MMI-Fuse, which apply the nest connection architecture [22] to extract the features of source images and an improved multi-attention model to fuse the features.
2. We proposed an information preservation-weighted channel-spatial attention module (ICS) to fuse features
more effectively and reasonably based on the information preservation degree.

(3). We performed efficient comparison experiments with other outstanding fusion methods on the widely used medical image dataset, Harvard dataset,1 where the results showed that the proposed method promotes the quality and efficiency of fusion of medical images.

The rest of the paper is structured as follows. We briefly introduce the related works in Section 2. The details of the proposed MMI-Fuse are presented in Section 3, and the experimental analyses are performed in Section 4. Finally, the conclusions are presented in Section 5.

II. RELATED WORKS
A. NEST CONNECTION
The theories and practices of skip connection have been known for a long time. He et al. [23] applied skip connections to avoid vanishing and exploding gradients, which made the CNN architecture deeper. However, Zhou et al. [22] suggested that the long-range skip connection in CNN architecture may lead to semantic gaps. Based on this observation, they first introduced the nest connection, Unet++, for medical image segmentation in 2018. UNet++ applied upsampling and short skip connections to replace long-range skip connections. The upsampling between different scale features released the semantic gap efficiently. Following this, Li et al. [24] introduced the nest connection into infrared and visible image fusion tasks and achieved noticeable performance. Consequently, we introduced nest architecture into the multimodal brain image fusion field and proposed a new fusion framework.

B. ATTENTION MECHANISM
The attention mechanism arose from the transformer model, which has achieved excellent performance in NLP over the years [25]. Many researchers have introduced this mechanism into CV due to the outstanding global and local information integration ability of attention. In image fusion, an attention mechanism was applied to design a fusion strategy. Vahishan et al. [26] proposed a spatiotemporal transformer fusion strategy with a spatial branch and an axial attention branch. The former branch captured local features, and the latter learned global-contexture features. Li et al. [24] designed a two-stage attention fusion strategy for deep feature fusion. This strategy included a spatial attention model and a channel attention model that was utilized to focus on local and global features, respectively. Inspired by these works, we designed a one-stage multi-attention model. The details are shown in the next section.

C. PERFORMANCE EVALUATION OF IMAGE FUSION
An efficient image fusion method should preserve the complementary information in the fused result as much as possible and make the fused image look more natural. To evaluate the performance of the fusion method, there are two ways: subjective and objective quality metrics [40]. In the medical image fusion field, the former way, including observations on distortion and spatial details, is simple and reliable. However, the evaluation results are affected by many factors, making the results unstable. The latter evaluates the fusion results through precise mathematical calculations and unified standards. Although it is complex, the results are stable and reliable. Consequently, many researchers are committed to proposing more universal objective quality metrics. Sengupta et al. [39] proposed three metrics for image fusion algorithms based on edge information. The authors utilized fractional order differentiation to obtain edge information. The edge information that was obtained using fractional order differentiation was used to estimate the values of three normalized weighted metrics for the fused image and corresponding source images.

III. PROPOSED METHOD
In this section, we showed the details of the proposed multimodal brain image fusion framework, MMI-fuse. This framework can be used to fuse either two medical images (such as pseudocolor SPECT or PET along with black and white MRI images), or three medical images (such as SPECT and MR-T2 as well as MR-Gad). We illustrate the overall structure of the proposed framework in Fig. 1. Specifically, this framework includes three parts: encoder blocks, the fusion strategy, and decoder blocks.

First, the source images are preprocessed by a head block to increase the channels from 3 to 64. Then, these 64 channel features are fed into four encoder blocks to extract features. The extracted features from the source images are then fused by the ICS block. Then, the fused features are fed into the decoder blocks to reconstruct the fused image. The details of the three parts will be introduced in the following.

A. AUTOENCODER-DECODER NETWORK
The main goal of the autoencoder-decoder network (AED) is to produce the features for the fusion module, information preservation weighted channel spatial attention (ICS), as shown in Fig. 1. In the encoder, four encoder blocks are stacked to extract features with four different scales from the input image. Each encoder block includes two convolutional layers to downsample the features. Similarly, the decoder block also consists of two convolutional layers. The only difference in the decoder is that, the dropout layer is replaced with the ReLu layer. Each decoder block receives the features of this layer and the upsampled features of the next layer at the same time. We have shown the details of the encoder block and decoder block in Fig. 2. The convolution kernel size of Conv layer 1 is 3, and that of Conv layer 2 is 1. The channels of the features are downsampled to half in Conv layer 1 and upsampled to the predefined output channel in Conv layer 2.

In the training stage, inspired by DenseFuse [27], the ICS block is separated from the main framework. According to the characteristics of PET and MRI images, the loss function

1http://www.med. harvard.edu/AANLIB/
we applied includes two parts.

$$L = \alpha L_{\text{spectral}} + \beta L_{\text{structure}}.$$  \hspace{1cm} (1)

where the spectral loss, $L_{\text{spectral}}$, focuses on the color information of PET and SPECT images and the structure loss, $L_{\text{structure}}$, focuses on the texture information of MRI images. $\alpha$ and $\beta$ are the weighting factors for the spectral and structure loss, respectively. The mathematical expression of the two loss functions is written below.

$$L_{\text{spectral}} = \frac{1}{H \times W} \| I_i - I_f \|_F^2.$$  \hspace{1cm} (2)

where $H$ and $W$ are the height and width of the source images, respectively. $I_i$ and $I_f$ are the input original pseudocolor image and fused image. $\| \cdot \|_F$ is the Frobenius norm.

$$L_{\text{structure}} = 1 - \text{SSIM} (O, I).$$  \hspace{1cm} (3)

where SSIM represents the structure similarity measure [28]. $I_i$ and $O$ denote the input images and output images, respectively. Considering the difficulty of collecting a medical image dataset, we trained the autoencoder-decoder network on COCO2014 [29]. During the training time, we set the parameters $\alpha$ as 1 and $\beta$ as 100. Then, the fusion features are restructured to obtain the fusion image.
B. FUSION STRATEGY

The attention mechanism plays a pivotal role in the human vision system. The vision signals are divided into several channels as they are forwarded in the brain. Then, the attention system of the brain filters out the key channels from these input signals according to the demands of the task while allocating less attention to other unimportant channels. Sanghyun Woo et al. [30] proposed a simple but efficient attention module to boost the performance of CNNs, to focus on the information of features of the channel and spatial axes. Inspired by this, we designed an attention module for medical image fusion, the information preservation weighted channel spatial attention (ICS) module, which carries out attention operations simultaneously and fuses the output of the channel and spatial branches. Fig. 3 illustrates the structure of the ICS in detail. The structure mainly includes three parts: channel attention, spatial attention, and information preservation weighting (IPW).

1) CHANNEL ATTENTION MODULE

This module focuses on the inter-channel relationship information while fusing the two input features. We first applied a convolution on the input features. Then, global pooling is utilized to compute the initial global weight \( W_c(I_k) \). The mathematical expression is shown in Eq. (4)

\[
W_c(I_k) = G(\Phi_i(I_k)).
\]

where \( \Phi_i \in \mathbb{R}^{C_i \times W_i \times H_i} \), \( W_c(I_k) \in \mathbb{R}^{C_i \times 1 \times 1} \), and \( I_k, (k \in \{1, 2\}) \) represent the four scale features extracted by the encoder and the two source input images, respectively. \( G(\cdot) \) denotes the global pooling operation. We choose the average operator to average all of the channels. The pooling operation is performed channel-wise. That is, the pooling kernel size is set the same as the \( w \) and \( h \) dimensions of the input feature. In Fig. 3, balance strategy 1 is a weighted mean operation that performs on \( W_c(I_k) \). The \( W_c(I_k) \) after weighted mean was repeated as the same size with input features in the \( w \) and \( h \) dimensions. Then, the output fusion feature of the channel attention module \( \Phi_{cl}^f \) is calculated by Eq. (5),

\[
\Phi_{cl}^f = \frac{W_c(I_1)}{\sum_{k=1}^{2} W_c(I_k) + \varepsilon} \cdot \Phi_i(I_1) + \frac{W_c(I_2)}{\sum_{k=1}^{2} W_c(I_k) + \varepsilon} \cdot \Phi_i(I_2). \tag{5}
\]

where \( \varepsilon = 0.0001 \) is used to avoid division by zero.

2) SPATIAL ATTENTION MODULE

This module focuses on the interspatial relationship information when the two source image features are fused. Different from the channel attention module, we use the mean value on the channel dimension to explore the interspatial information. The initial spatial weight \( W_s^I(x, y) \) is calculated using Eq. (6),

\[
W_s^I(x, y) = \frac{\sum_{i=1}^{C} \Phi_i^s(x, y)}{C}, \tag{6}
\]

where \( W_s^I(x, y) \in \mathbb{R}^{1 \times W_i \times H_i} \), and \( C \) denotes the channel number. The meaning of the parameters in Eq. (6) are the...
same as in Eq. (4). Then, we introduce balance strategy 2 to compute the spatial attention weight for the two source image features. The output of the spatial attention model $\Phi_{si}^f$ is calculated by Eq. (7),

$$
\Phi_{si}^f = \frac{e^{W_{i1}(x,y)}}{\sum_{k=1}^{2} e^{W_{ik}(x,y)} + \epsilon} \cdot \Phi_{i}(I_1) + \frac{e^{W_{i2}(x,y)}}{\sum_{k=1}^{2} e^{W_{ik}(x,y)} + \epsilon} \cdot \Phi_{i}(I_2).
$$

(7)

where $\epsilon = 0.0001$ is used to avoid division by zero.

3) INFORMATION PRESERVATION WEIGHTING

Gradients in features are easy to compute and store [31]. To fuse the output features from the channel attention and spatial attention models, we introduce an information preservation weighting strategy that evaluates the information content according to the gradients of these two features. The information preservation degree of the output features is defined in Eq. (8),

$$
G_{i_k} = \frac{1}{H_i W_i C_i} \sum_{c=1}^{C_i} (\nabla \Phi_{i}^c(O_k))^2.
$$

(8)

where $\Phi_{i}(O_k) \in \mathbb{R}^{C_i \times W_i \times H_i}$ denotes the output feature of the $k$-th input source image from the channel or spatial attention module. $H_i$, $W_i$, and $C_i$ are the height, width and channel, respectively. $\nabla$ is the Laplacian operator. We fuse the features processed by the channel and spatial attention modules with the adaptive information preservation weight. We refer to $W_{ik}$ to denote this adaptive weight. The mathematical expression of $W_{ik}$ is written as Eq. (9),

$$
[W_{1,}, W_{2,}] = \text{Softmax} \left(G_{i_1}, G_{i_2}\right).
$$

(9)

where we applied the Softmax function to set $W_{1,}, W_{2,}$ between 0 and 1. Then, the two weights are used to fuse features processed by attention modules.

IV. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTAL CONDITIONS

1) SOURCE IMAGES

We collected 106 pairs of multimodal medical images from the widely accepted Harvard dataset released by Harvard Medical School. These image pairs include four brain diseases: AISD, Alzheimer’s disease, glioma, and Huntington’s disease. Among these images, 20 pairs of proton density-weighted MRI (MRI-PD) and T2-weighted MRI (MRI-T2) images, 50 pairs of MRI (proton density-weighted and T2-weighted) and SPECT images, 16 pairs of gadolinium-weighted MRI and SPECT images, and 20 pairs of MRI (T1-weighted and T2-weighted) and fluorodeoxyglucose PET images were included in our collection. All source images were 256 × 256 pixels and were well registered. Considering that the volume of data was small, we extended the collected data. We adopted two data enhancement methods: flipping and clipping. First, we flipped all of the images twice, from right to left and from the top to down. This operation increases the number of image pairs to 318. Then, we clipped the size of the images to 64 × 64; that is, from every 256 × 256 image, four sub-images are generated. Finally, we obtained 1590 pairs of images.

2) COMPARISON EXPERIMENT

To illustrate the performance of the proposed method, we compared it with seven recent approaches. These approaches include the LLF-IOI [5], CNNs [32], PAPCNN [7], CTD-SR [33], MRNDM [34], TA-cGAN [21], and U2Fuse [31].

3) EVALUATION METRICS

Initially, we conducted a subjective comparison among all methods, but this is insufficient to evaluate the performance of our fusion approach. Consequently, we applied five metrics to analyze the quality of the fusion images. Specifically, these metrics include entropy (EN) [35], the gradient-based index ($Q_{AB/F}$) [36], the visual information fidelity for fusion (VIF) [37], the union structural similarity measurement ($SSIM_u$), and the standard deviation (SD). Considering that there are two source images that can be seen as the reference images, the structural similarity measurement in this paper was modified to $SSIM_u$. The expression of $SSIM_u$ is shown in Eq. (10),

$$
SSIM_u(f) = \frac{(SSIM(f, I_1) + SSIM(f, I_2))}{2}.
$$

(10)

where $SSIM(f, I_1)$ and $I_2$ denote the standard version of the structural similarity measurement, the fused image and two input images, respectively.

B. FUSION OF PSEUDOCOLOR AND GRAYSCALE IMAGES

SPECT and PET imaging generate pseudocolor images in RGB format with 3 channels, whereas MR imaging generates grayscale images with 1 channel. This kind of incompatibility on the channel numbers limits the fusion of multimodal medical images in one step. To fuse input images with
FIGURE 5. Three sets of T1-weighted MRI and fluorodeoxyglucose PET image fusion results. The first column is the input source images. $a_i$, $b_i$, $c_i$, $d_i$, $e_i$, $f_i$, $g_i$, and $h_i$ represent the results of CNNs, CTD-SR, LLF-IOI, NDM, PA-PCNN, TA-cGAN, U2fuse, and the proposed method, respectively.

different number of channels, we first converted the RGB images into the YCbCr space. The YCbCr color space encodes a color image with an illumination component (Y), blue hue component (Cb), and red hue component (Cr). Then, the Y part of the YCbCr space was fused with the grayscale source image. Finally, the fused image can
be converted to RGB space through an inverse conversion function. Fig. 4 shows the processing procedure of the RGB and grayscale source images in our framework. The blue lines denote the flow of the image data.

C. SUBJECTIVE VISUAL EVALUATION

We analyzed the results of the proposed method along with the results of other widely used methods. The fused medical images consisted of five different
FIGURE 7. Three sets of SPECT-Tc and MRI-PD image fusion results. The first column is the input source images. $a_i$, $b_i$, $c_i$, $d_i$, $e_i$, $f_i$, $g_i$, and $h_i$ represent the results of CNNs, CTD-SR, LLF-IOI, NDM, PA-PCNN, TA-cGAN, U2Fuse, and the proposed method, respectively.

combinations, including MRI-PD and MRI-T2, MRI-PD and SPECT-Tc, gadolinium-based MRI and SPECT-Tc, gadolinium-based MRI and SPECT-Tl, and MRI and fluorodeoxyglucose PET.
Fig. 8 shows the results of T1-weighted MRI and fluorodeoxyglucose PET image fusion for visual comparison. It is obvious that the fusion results of CTD-SR, PA-PCNN, and TA-cGAN lost the structural information in the MRI-T1 image \((b_i, e_i, \text{ and } f_i, i \in (1, 2, 3))\). The CNNs, NDM, and LLF-IOI perform better in structure information preservation.
TABLE 1. Objective quality of different methods on five metrics.

| Images          | Metrics | CNNs | CTD-SR | LLF-IOI | NDM | PA-PCNN | TA-cGAN | U2fuse | Proposed |
|-----------------|---------|------|--------|---------|-----|---------|---------|--------|----------|
| MRI-T1 and PET  | EN      | 4.875 | 4.3049 | 4.7882  |     | 4.9965  | 5.3199  | 4.6349  | 5.1410   | 5.0556   |
| MRI-PD          | SD      | 61.776 | 58.2150 | 61.1946 |     | 53.3267 | 63.8392 | 62.7524 | 57.8073 | 76.3199 |
| MRI-T2          | Q_{ARI} | 0.6478 | 0.5139 | 0.5383 |     | 0.6884 | 0.5411 | 0.5716 | 0.3297 | 0.6781 |
| MRI-PD and SPECT-Ta | VIFF | 0.6807 | 0.5569 | 0.5405 |     | 0.6427 | 0.7153 | 0.5817 | 0.3588 | 0.6286 |
| MRI-PD and SPECT-Ta | SSIM_m | 0.6437 | 0.5196 | 0.5361 |     | 0.5824 | 0.5984 | 0.6083 | 0.4529 | 0.6191 |

However, the structure information is ambiguous where the pseudocolor is noticeable (a_i, c_i, and d_i, i ∈ (1, 2, 3)). U2fuse performs better in both structural information and pseudocolor information preservation in the intracranial area but with low contrast in the skull and maxillofacial area (g_i, i ∈ (1, 2, 3)). Compared with these methods, the proposed method realizes accurate registration between the anatomical structures and functional metabolism. Moreover, the anatomical structures in gray matter, white matter, sulci and gyri are clearer than the existing methods mentioned above. Reflecting the corresponding functional metabolism on the clear anatomical structure is more conducive to comprehensive clinical judgment (h_i, i ∈ (1, 2, 3)).

Fig. 6 illustrates the visual comparison results of MRI-T2 and MRI-PD image fusion. The results of CTD-SR, see (b_i), retain less structural information in both MRI-T2 and MRI-PD images. The results of U2fuse are better in preserving the details of brain tissue (g_i). Compared with CNNs, LLF-IOI, PA-PCNN, NDM, and TA-cGAN, the proposed method can obtain almost the same performance in less time, as shown in TABLE 2. However, the performance of U2fuse exceeds the proposed method in terms of contrast (g_i).

Fig. 7 presents the fusion results of MRI-PD and SPECT-Ta images. The results of LLF-IOI, CTD-SR, and PA-PCNN contain more pseudo color information in SPECT images (b_i, c_i and e_i), which leads to the structural information in MRI-PD being covered well. The results of U2fuse are low contrast, which makes the outline of eyes, scalp fit and skull blurry (g_i). The results of the proposed method (h_i), show that both the pseudocolor information and structure details are taken into account. Although the visual quality of the proposed method is similar to CNNs, NDM, and TA-cGAN, our method takes less time.

Fig. 8 exhibits the fusion results of MRI-Gad and SPECT-Ta images. It is obvious that the gyrus formations are covered by the pseudocolors in the fusion results of CTD-SR, LLF-IOI, and NDM. In contrast, the results of the proposed method are sharper than other methods in terms of the gyrus details. The last two rows show the fusion results of MRI-Gad and SPECT-Ta.

D. OBJECTIVE QUALITY

The objective quality of different methods is presented using five metrics. The average values of five metrics for each fusion combination. For each metric, we showed the highest value among the eight

| Images          | CNNs | CTD-SR | LLF-IOI | NDM | PA-PCNN | TA-cGAN | U2fuse | Proposed |
|-----------------|------|--------|---------|-----|---------|---------|--------|----------|
| MRI-Gad and SPECT-Ta | SD   | 64.904 | 63.1006 | 66.2586 |     | 66.7228 | 68.6817 | 68.2139 | 42.1398 | 71.7897 |
| MRI-PD and SPECT-Ta | VIFF | 1.531 | 1.8332 | 1.6837 |     | 1.7574 | 2.1115 | 2.3566 | 1.2711 | 1.8729 |
| MRI-PD and SPECT-Ta | SSIM_m | 0.5366 | 0.6119 | 0.6097 |     | 0.6218 | 0.6537 | 0.678 | 0.6187 | 0.6593 |
| MRI-Gad and SPECT-Ta | SD   | 56.554 | 61.4893 | 52.8432 |     | 48.6629 | 65.0624 | 64.4563 | 48.7588 | 54.6549 |
| MRI-PD and SPECT-Ta | VIFF | 0.0728 | 0.0704 | 0.0761 |     | 0.0659 | 0.0704 | 0.0819 | 0.0753 | 0.0779 |
| MRI-PD and SPECT-Ta | SSIM_m | 0.5177 | 0.5449 | 0.4992 |     | 0.5549 | 0.5713 | 0.5345 | 0.4926 | 0.5821 |

However, the structure information is ambiguous where the pseudocolor is noticeable (a_i, c_i, and d_i, i ∈ (1, 2, 3)). U2fuse performs better in both structural information and pseudocolor information preservation in the intracranial area but with low contrast in the skull and maxillofacial area (g_i, i ∈ (1, 2, 3)). Compared with these methods, the proposed method realizes accurate registration between the anatomical structures and functional metabolism. Moreover, the anatomical structures in gray matter, white matter, sulci and gyri are clearer than the existing methods mentioned above. Reflecting the corresponding functional metabolism on the clear anatomical structure is more conducive to comprehensive clinical judgment (h_i, i ∈ (1, 2, 3)).

Fig. 6 illustrates the visual comparison results of MRI-T2 and MRI-PD image fusion. The results of CTD-SR, see (b_i), retain less structural information in both MRI-T2 and MRI-PD images. The results of U2fuse are better in preserving the details of brain tissue (g_i). Compared with CNNs, LLF-IOI, PA-PCNN, NDM, and TA-cGAN, the proposed method can obtain almost the same performance in less time, as shown in TABLE 2. However, the performance of U2fuse exceeds the proposed method in terms of contrast (g_i).

Fig. 7 presents the fusion results of MRI-PD and SPECT-Ta images. The results of LLF-IOI, CTD-SR, and PA-PCNN contain more pseudo color information in SPECT images (b_i, c_i and e_i), which leads to the structural information in MRI-PD being covered well. The results of U2fuse are low contrast, which makes the outline of eyes, scalp fit and skull blurry (g_i). The results of the proposed method (h_i), show that both the pseudocolor information and structure details are taken into account. Although the visual quality of the proposed method is similar to CNNs, NDM, and TA-cGAN, our method takes less time.

Fig. 8 exhibits the fusion results of MRI-Gad and SPECT-Ta images. It is obvious that the gyrus formations are covered by the pseudocolors in the fusion results of CTD-SR, LLF-IOI, and NDM. In contrast, the results of the proposed method are sharper than other methods in terms of the gyrus details. The last two rows show the fusion results of MRI-Gad and SPECT-Ta.

The color information of SPECT-Ta is represented less in the fusion results of CNNs, CTD-SR, LLF-IOI, NDM, and TA-cGAN, so the colors look shallow. The PA-PCNN achieves the best visual performance.
methods, including the proposed method, with red boldface. It is easy to see that the proposed method reported best performance for 8 times, which means that the proposed method ranks first among the 8 methods. The TA-cGAN is the second having the top score for 6 times, and the PA-PCNN and the U2fuse are in the third place having the top score for 4 times. In particular, the proposed method performs best on SD and SSIMu. The performance of the TA-cGAN in the five metrics is outstanding overall. U2fuse always takes the top score in terms of EN. There is no prominently advantageous score among the rest of the methods.

**E. EFFICIENCY COMPARISON**

TABLE 2 reports the run times of the 8 methods. We took 10 pairs of MRI-T1 and PET images for evaluation. The best score is shown in boldface with red. The data in TABLE 2 represent the total time spent by the corresponding method to fuse 10 pairs of images. We can see that the proposed method takes less time for fusing 10 pairs of images (only 6.97 seconds), while it also achieves the best fusion results.

**F. ABLATION STUDY**

1) HYPERPARAMETERS $\alpha$ AND $\beta$

In our study, the hyperparameters $\alpha$ and $\beta$ in Eqn. (1) are self-defined. To explore the best $\alpha$ and $\beta$ combination, we set $\alpha$ as 0.1, 0.5, and 1 and $\beta$ as 1, 10, 100, and 1000 for the ablation study.

In Fig. 9, we show the fused results of MRI-PD and SPECT images with different $\alpha$ and $\beta$ combinations as an example. The four columns from left to right represent that $\beta$ equals 1, 10, 100, 1000, and the three columns from top to bottom represent that $\alpha$ equals 0.1, 0.5, 1, respectively.
Subjectively, it is obvious that the performance of MMI-fuse is unsatisfactory when the value of $\alpha$ is less than 0.5 and $\beta$ is less than 100. On the other hand, when $\alpha$ equals 1 and $\beta$ equals 1000, the increase in local brightness causes the structural information to be covered.

From the results shown above, assigning a larger $\alpha$ and $\beta$ is good for retaining structure information. To make better choices, we compared the objective qualities of different combinations of the hyperparameters. TABLE 3 shows the results of the objective metrics. The data indicates that although combination $\alpha = 1$ and $\beta = 100$ is inferior to combination $\alpha = 1$ and $\beta = 1000$ in terms of some individual indicators, such as SD and $Q_{AB/F}$, the overall results of the former are better than those of the latter.

2) ATTENTION MODULE IN ICS

The fusion strategy plays a crucial role in the proposed method. This section analyzes the influence of different attention module combinations on the fusion results. In Fig. 10, we exhibit several fusion results for visual analysis. The channel attention module can retain more structural information while decreasing the contrast between tissues with different densities (skull and brain tissues). Spatial attention can obtain good contrast but blur the sulcus information. Channel and spatial attention improved the blur effect of the sulcus but ignored the contrast. In contrast, our proposed ICS module not only improved the information of the sulci but also improved the contrast of tissues with different density.

We show the objective metrics in TABLE 4. From the results, it is seen that the proposed ICS module reports the best performance in EN, SD, VIFF, and SSIM. Compared with the single attention module, the multi-attention module with information preservation weight achieved more remarkable results.

V. CONCLUSION

In this paper, we proposed a multimodal brain image fusion framework based on an improved multichannel attention model, MMI-fuse. This framework utilizes a nested architecture for feature extraction and image restructuring. In terms of the fusion strategy, we proposed a novel information preservation weighted channel and spatial attention module, which adaptively fuses the features processed by the channels and spatial attention modules. To validate the performance of MMI-fuse, we collected 1590 pairs of multimodal medical images from the Harvard dataset to conduct extensive experiments. Seven methods, including CNNs, CTD-SR, LLF-101, NDM, PA-PCNN, TA-cGAN, and U2fuse, and five metrics, EN, SD, $Q_{AB/F}$, VIFF, and SSIM, were selected for validation. The results show that while the proposed method achieves the best performance on some of the metrics, it costs the least time among all of these seven methods. Although the proposed MMI-fuse method achieved satisfactory results, there are still many scores that could be improved. For instance, our method has not achieved significant results in MRI and SPECT image fusion. We found that the feature extractor is another important component affecting the performance of the fusion framework. In addition, this work focuses on the fusion of multimodal brain images and lacks generalization to other multimodal image fusions. Consequently, we will focus on the improvement of the encoder and decoder to extract features and reconstruct fused images and perform more horizontal experiments to increase the generalization ability of the model. Finally, the pseudocolor fidelity of MMI-fuse needs to be improved further.

REFERENCES

[1] A. Seal, D. Bhattacharjee, M. Nasipuri, D. Rodríguez-Esparragón, E. Menasalvas, and C. Gonzalo-Martin, “PET-CT image fusion using random forest and à-trous wavelet transform,” Int. J. Numer. Methods Biomed. Eng., vol. 34, no. 3, p. e2933, Mar. 2018, doi: 10.1002/cnm.2933.

[2] S. Polinati, D. P. Bavirisetti, K. N. V. P. S. Rajesh, and R. Dhuli, “Multimodal medical image fusion based on content-based decomposition and PCA-sigmoid,” Current Med. Imag., vol. 18, no. 5, pp. 546–562, Apr. 2022, doi: 10.2174/1573405617666211004114726.
A nested

Z. Zhou, M. Siddiquie, N. Tjahbijaks and J. Liang, “UNET++: A nested U-Net architecture for medical image segmentation,” in Procedia Computer Science, vol. 36, pp. 3–11, 2014, doi: 10.1016/j.procs.2014.09.004.

K. He, X. Zhang, S. Ren and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

H. Li, X.-J. Wu and T. Durrani, “NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention modules,” IEEE Trans. Instrum. Meas., vol. 69, no. 12, pp. 9645–9656, Dec. 2020, doi: 10.1109/TIM.2020.3005230.

N. Kitaev, L. Kaiser and A. Levskaya, “Reformer: The efficient attention transformer,” 2020, arXiv:2001.00451.

V. S. Vishuban, J. M. J. Valanarasu, P. Oza and V. M. Patel, “Image fusion transformer,” Aug. 2021, arXiv:2107.09011.

H. Li and X.-J. Wu, “DenseFuse: A fusion approach to infrared and visible images,” IEEE Trans. Image Process., vol. 28, no. 5, pp. 2614–2624, May 2019, doi: 10.1109/TIP.2018.2878342.

Z. Wang, A. C. Rovik, 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, doi: 10.1109/TIP.2003.818961.

T. Y. Lin, M. Maire, S. Belongie, J. Hays and C. L. Zitnick, “Microsoft COCO: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., Zurich, Switzerland: Springer, Feb. 2015, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

S. Woo, J. Park, J.-Y. Lee and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., Jul. 2018, pp. 3–19, doi: 10.1007/978-3-030-01234-2_1.

H. Xu, J. Ma, J. Jiang, X. Guo and H. Ling, “U2Fusion: A unified unsupervised image fusion network,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 1, pp. 502–518, Jan. 2022, doi: 10.1109/TPAMI.2020.3011042.

Y. Liu Li, Chen, J. Cheng and H. Peng, “A medical image fusion method based on convolutional neural networks,” in Proc. 20th Int. Conf. Inf. Fusion (Fusion), Jul. 2017, pp. 1–7, doi: 10.23919/ICIF.2017.809769.

Z. Zhu, H. Yin, Y. Chai, Y. Li and G. Qi, “A novel multi-modality image fusion method based on image decomposition and sparse representation,” Int. J. Comput. Assist. Radiol. Surg., vol. 4, pp. 516–529, Mar. 2019, doi: 10.1007/s11548-017-1377-1.

S. Liu, Y. Song, V. S. Sheng, C. Xu, C. Maere, K. Xue and K. Yang, “MRI and PET image fusion using the nonparametric density model and the theory of variable-weight,” Comput. Methods Programs Biomed., vol. 175, pp. 73–82, Jul. 2019, doi: 10.1016/j.cmpb.2019.04.010.

W. J. Roberts, J. A. A. Van and F. Ahmed, “Assessment of image fusion procedures using entropy, image quality, and multispectral classification,” Proc. SPIE, vol. 120, no. 1, pp. 1–7, May 2008, doi: 10.1117/1.2945910.

C. S. Xydeas and V. Petrovic, “Objective image fusion performance measure,” Mil. Technol. Courier, vol. 56, no. 4, pp. 181–193, 2000, doi: 10.5937/vojtehg0802181B.

Z. Han, Y. Cai, Y. Cao and X. Xu, “A new image fusion performance measure based on visual information fidelity,” IEEE Trans. Image Process., vol. 14, no. 2, pp. 127–135, Apr. 2013, doi: 10.1109/TIP.2011.2180097.

H. Li, Y. Wang, Z. Yang, R. Wang, X. Li and D. Tao, “Discriminative dictionary learning-based multiple component decomposition for detail-preserving noisy image fusion,” IEEE Trans. Instrum. Meas., vol. 69, no. 4, pp. 1082–1102, Apr. 2020, doi: 10.1109/TIM.2019.2912239.

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. 86258–86278, 2020, doi: 10.1109/ACCESS.2020.2993607.

B. Meher, S. Agrawal, R. Panda and A. Abraham, “A survey on region based image fusion methods,” Inf. Fusion, vol. 48, pp. 119–132, Aug. 2019, doi: 10.1016/j.inffus.2018.07.010.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, “Backpropagation applied to handwritten zip code recognition,” Comput., vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: 10.1109/ECO.1989.451451.

Y. Li, J. Zhao, Z. Lv and Z. Pan, “Multimodal medical supervised image fusion method by CNN,” Frontiers Neurosci., vol. 15, p. 303, Jun. 2021, doi: 10.3389/fnins.2021.638976.

K. Guo, X. Li, X. Hu, J. Liu and T. Fan, “Hahn-PCNN-CNN: An end-to-end multi-modal brain medical image fusion framework useful for clinical diagnosis,” BMC Med. Imag., vol. 21, no. 1, p. 111, Jul. 2021, doi: 10.1186/s12880-021-00642-z.
ZHENGHE SHI received the B.Eng. degree from the School of Electrical Engineering Automation, Hunan University of Arts and Science, Changde, China, from 2016 to 2020. He is currently pursuing the master’s degree in instrument science and technology with the School of Mechanical Engineering, Xi’an University of Science and Technology, Xi’an, China. His research interests include driverless image processing and object detection.

CHUANWEI ZHANG received the B.Eng. and M.A.Eng. degrees in mechatronic engineering from the Xi’an University of Science and Technology, Xi’an, China, in 1998 and 2001, respectively, and the Ph.D. degree in mechatronic engineering from Xi’an Jiaotong University, China, in 2006. From 2003 to 2007, he was a Lecturer with the Mechanical Engineering Department, Xi’an University of Science and Technology, where he has been a Professor with the Mechanical Engineering Department, since 2007. He is the author of three books, more than 50 articles, and more than 50 inventions. His research interests include intelligent network vehicles and autonomous navigation in complex environments.

DAN YE received the B.S. degree in clinical medicine from Xi’an Jiaotong University, Xi’an, China, from 2015 to 2020, where she is currently pursuing the master’s degree in dermatology and venereology with the Second Affiliated Hospital, School of Medicine, receiving the standardized residents training.

PEILIN QIN received the B.Eng. degree in mechatronic engineering from the Xi’an University of Science and Technology, Xi’an, China, in 2017, and the joint M.A.Eng. degree in mechanical engineering from the Xi’an University of Science and Technology and Brunel University London, London, U.K., in 2020. He is currently pursuing the Ph.D. degree in autonomous robotics with the Xi’an University of Science and Technology. His research interests include computer vision, sensor fusion, and deep neural networks for autonomous robots.

RUI ZHOU received the B.Eng. degree in mechanical engineering from the Xi’an University of Science and Technology, Xi’an, China, in 2019, where she is currently pursuing the M.A.Eng. degree in autonomous robotics. Her research interests include sensor fusion, path planning, and intelligent navigation for autonomous cars.

LEI LEI received the bachelor’s degree in mechanical and electronic engineering from the Xi’an University of Science and Technology, in 2020, where he is currently pursuing the master’s degree in unmanned driving. His research interests include simultaneous localization and mapping and sensor fusion.

***