GeSeNet: A General Semantic-Guided Network With Couple Mask Ensemble for Medical Image Fusion

Jiawei Li, Jinyuan Liu, *Member, IEEE*, Shihua Zhou, *Member, IEEE*, Qiang Zhang, *Senior Member, IEEE*, and Nikola K. Kasabov, *Life Fellow, IEEE*

***Abstract*— At present, multimodal medical image fusion technology has become an essential means for researchers and doctors to predict diseases and study pathology. Nevertheless, how to reserve more unique features from different modal source images on the premise of ensuring time efficiency is a tricky problem. To handle this issue, we propose a flexible semantic-guided architecture with a mask-optimized framework in an end-to-end manner, termed as GeSeNet. Specifically, a region mask module is devised to deepen the learning of important information while pruning redundant computation for reducing the runtime. An edge enhancement module and a global refinement module are presented to modify the extracted features for boosting the edge textures and adjusting overall visual performance. In addition, we introduce a semantic module that is cascaded with the proposed fusion network to deliver semantic information into our generated results. Sufficient qualitative and quantitative comparative experiments (i.e., MRI-CT, MRI-PET, and MRI-SPECT) are deployed between our proposed method and ten state-of-the-art methods, which shows our generated images lead the way. Moreover, we also conduct operational efficiency comparisons and ablation experiments to prove that our proposed method can perform excellently in the field of multimodal medical image fusion. The code is available at https://github.com/lok-18/GeSeNet.**

***Index Terms*— Image fusion, multimodal medical image, region mask, semantic information.**

Manuscript received 19 November 2022; revised 18 May 2023; accepted 2 July 2023. This work was supported in part by the 111 Project under Grant D23006; in part by the National Natural Science Foundation of China under Grant 62272079 and Grant 61972266; in part by the Liaoning Revitalization Talents Program under Grant XLYC2008017; in part by the Natural Science Foundation of Liaoning Province under Grant 2021-MS-344, Grant 2021-KF-11-03, and Grant 2022-KF-12-14; in part by the Postgraduate Education Reform Project of Liaoning Province under Grant LNYJG2022493; and in part by the Dalian Outstanding Young Science and Technology Talent Support Program under Grant 2022RJ08. *(Corresponding authors: Shihua Zhou; Qiang Zhang.)*

Jiawei Li and Shihua Zhou are with the Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering, Dalian University, Dalian 116622, China (e-mail: ljw19970218@163.com; zhoushihua@dlu.edu.cn).

Jinyuan Liu is with the School of Mechanical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: atlantis918@hotmail.com).

Qiang Zhang is with the School of Computer Science and Technology, Dalian University of Technology, Dalian 116024, China (e-mail: zhangq26@126.com).

Nikola K. Kasabov is with the Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland 1061, New Zealand, also with the Intelligent Systems Research Center, University of Ulster, BT48 7JL Londonderry, U.K., and also with the IICT, Bulgarian Academy of Sciences, 1000 Sofia, Bulgaria (e-mail: nkasabov@aut.ac.nz).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3293274.

Digital Object Identifier 10.1109/TNNLS.2023.3293274

I. Introduction

WITH the rapid development of medical imaging technology during the past decades, multimodal medical images have been widely applied in clinical diagnosis, medical research, and surgical navigation [1]. Due to the difference between imaging equipment and techniques, various kinds of medical images highlight different information (e.g., bone contours and the location of the tumor), which can be roughly separated into two categories, that is, structural medical images and functional medical images [2]. As a typical kind of structural medical image, magnetic resonance imaging (MRI) images perform soft-tissue structure information for doctors and researchers to study. Computed tomography (CT) images can provide an outline of bone structure and brain anatomical information clearly with high resolution. However, structural medical images such as MRI and CT images are insensitive to functional information in human metabolism [3].

As a representative in the field of functional medical images, positron emission tomography (PET) images play an important role [4], which characterizes metabolic function, blood flow, and some tumor information in brain tissue. In addition, single-photon emission CT (SPECT) images as another vital functional medical image can highlight tissue damage and organ information [5]. Nonetheless, functional medical images still suffer from low-resolution performance and disable to accurately display structural information. Therefore, combining advantages from different modality medical images and merging them into a single image can not only improve the visual effect and complementarity of images, but also help doctors improve the accuracy of clinical diagnosis and disease forecasting [6].

Traditional and deep-learning-based methods are widely utilized in existing multimodal image fusion methods. Regardless of traditional or deep-learning-based methods [7], [8], their common purpose is to extract practical features from different single-source images and generate vivid fused images through a designed fusion strategy or network model. In most traditional methods, feature fusion rules based on spatial [9] and transform domain transformation [10] are employed generally. These traditional methods generate a new fused image by

transforming specific regions and then reconstructing them together. However, drawing complicated fusion rules manually is an inevitable procedure in traditional methods, which makes the efficiency of the fusion process reduce. Furthermore, fusion results may appear as undesirable artifacts by using the same decomposition operation to handle source images of different modalities.

In recent years, to ameliorate the disadvantages of traditional methods, deep-learning-based approaches are introduced to conduct multimodal image fusion tasks. Researchers can avoid the complexity of handcrafted fusion rules via end-to-end models [11]. Moreover, different modules in architecture can correctly extract their unique features from multiple single-modal images. Nevertheless, there still exist several limitations: 1) semantic information is often ignored in the multimodal fusion task, so that some artifact halos may occur, debasing the quality of the generated results; 2) some existing deep-learning-based approaches increase network scales to improve the quality of fused images, which causes a large number of redundant computations and makes running time too long; and 3) owing to the inaccurate extraction of prominent features in each source image, it is a great challenge to perform well in some texture details of fused images.

To alleviate these above-mentioned limitations, in this article, we proposed a flexible semantic-guided framework with mask-optimized models in an end-to-end manner for fusing multimodal medical images, called GeSeNet. Specifically, we concatenate our proposed fusion network with a pretrained semantic module. With the help of the semantic module, the semantic information of our fusion results can be increased significantly. For combining the extracted semantic information with fused images more realistically, we proposed an edge enhancement module and a corresponding edge loss function to cooperate with our network to highlight edge textures. The region mask module is introduced to identify "principal" and "redundant" regions in source images. As the flexible division of different regions, our proposed method can reduce superfluous computing and promote operational efficiency. Besides, we employ a global refinement module to optimize features extracted from the edge and mask modules, which can recover more textural details and achieve fusion results with fine visual effects. Fig. 1 compares U2Fusion [12] and EMFusion [13] with our proposed method through a set of MRI-CT, MRI-PET, and MRI-SPECT fused results. Noticeably, our method leads the way.

In short, we summarize our proposed work as the following four contributions.

1) We devise a novel fusion network called GeSeNet for fusing multimodal medical images, including MRI-CT, MRI-PET, and MRI-SPECT pairs. Different from previous deep-learning methods, we introduce a pretrained semantic module and a newly designed semantic loss function to cascade with our fusion network, so that some missing details of fused images can be complemented during the training phase.
2) We propose an edge enhancement module with a gradient filter and formulate its edge loss function in our proposed network. By guiding the network training via back-propagation, the enhanced-edge features can be extracted as a prior condition to reduce the appearance of edge artifacts and achieve realistic fused results.
3) Through the discrimination of the region mask module, our proposed method can intensify the representation of significant features while skipping the redundant computation to complete the whole fusion process in less time. Moreover, to get higher-performance fused results, we initiate a global refinement module for revising extracted details simultaneously from the edge enhancement and region mask modules.

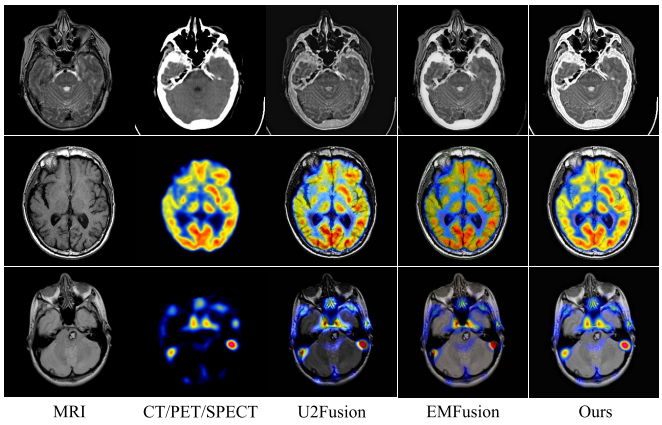

Fig. 1. Schematic illustration of our proposed method. Clearly, compared with U2Fusion and EMFusion, the proposed method provides more attractive details, edge textures, and faithful color on different multimodal medical image fusion tasks.

The remainder of this article is established as follows. Section II summarizes the related works of multimodal medical image fusion. Section III gives a detailed interpretation of the proposed method, including the overall framework, the edge enhancement module, the region mask module, the global refinement module, and the loss function. In Section IV, extensive qualitative and quantitative experiments are conducted to verify the advantage of our proposed method. Furthermore, we perform ablation experiments to analyze the effect of each module. Finally, the conclusion is given in Section V.

II. Related Works

In this section, we review previous studies about multimodal medical image fusion from model-driven methods, data-driven deep-learning methods, and model-driven deep-learning methods.

A. Model-Driven Methods

In the past, traditional methods use models to drive for fusion. The multiscale transform-based (MST) approach, for example, wavelet transform [14], pyramid transform [15], and subspace transform [16], is the most commonly employed in traditional methods. In MST approaches, researchers often transform source images into a mutable matrix, fuse related parameters, and implement inverse matrix transformation to complete the fusion process. Moreover, sparse

representation-based methods [17], salient feature-based methods [18], and so on are also widely applied in multimodal medical image fusion.

Specifically, the wavelet transform-based methods can be roughly divided into discrete wavelet [19], stationary wavelet [20], lifting wavelet [21], and so on. As a representative, Cheng et al. [22] proposed an innovative architecture based on wavelet transform to achieve the goal of fusing CT images with PET images, which can exactly detect pathological changes. Bhavana and Krishnappa [23] first employed Gaussian filters to preprocess source images and then used the discrete wavelet to enhance the performance of fusion results. Quantitative indicators, that is, average gradient and spectral discrepancy, achieve high marks by using this method. Ganasala and Prasad [10] introduced a novel approach based on stationary wavelet transformation and texture energy measures to solve issues of poor contrast and low computing ability.

Laplacian pyramid transform-based methods also perform extensively in multimodal medical image fusion. Sahu et al. [24] utilized the Laplacian pyramid with discrete cosine transform (DCT) to decompose source images as different low-pass-filtered patches. The quality of fused images is positively related to the number of levels in the pyramid. He et al. [25] integrated the advantages of intensity–hue–saturation (IHS) transform and principal component analysis (PCA) to improve the performance of fused images. Krishn et al. [26] used PCA to maximize the spatial resolution on the decomposed coefficients.

B. Data-Driven and Model-Driven Deep-Learning Methods

During the last decade, it has become rapidly popular that scholars use deep-learning-based methods to solve multimodal medical image fusion, which can be divided into data-driven and model-driven approaches [27], [28], [29], [30].

As a representative, Singh and Anand [31] presented a novel method with a two-scale l_1-l_0 hybrid layer decomposition scheme to avoid artifacts and noise on the feature level. With this approach, they could fuse source images in the decomposed base and detail layer. Liu et al. [32] introduced a Siamese convolutional network to obtain a weight map that contains the pixel activity information from inputs. A local similarity strategy was employed to regulate the fusion mode. Song et al. [33] proposed a multiscale DenseNet called MSD-Net through an encoder–decoder model, which used three different filters to extract features.

In some unified fusion frameworks, multimodal medical image fusion has become an important branch to reveal its comprehensiveness. Zhang et al. [34] presented a novel method with excellent generalization ability to improve perceptual information in fused images. Xu et al. [12] proposed an adaptive retention mechanism to conduct multimodal (i.e., infrared and visible images and medical images), multiexposure, and multifocus image fusion. Liu et al. [35] mentioned a bilevel optimization paradigm for multimodal image fusion, which used a formulaic decomposition method to complete fusion processing between two modalities.

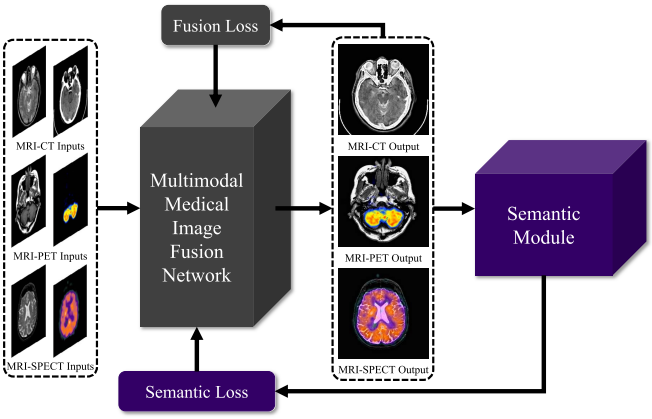

Fig. 2. Overall framework of the proposed method. Source images are first input into a multimodal medical image fusion network to generate original fusion images and then fed into a semantic module to extract semantic information. Finally, by passing the semantic information to the fusion network for reprocessing, we can obtain well-performed fusion results.

Moreover, pulse coupled neural network (PCNN)-based methods are also very active in the field of multimodal medical image fusion. Wang and Ma [36] proposed a novel multichannel model, that is, m-PCNN, to deal with different models of medical images for the first time. Xu et al. [37] introduced the adaptive PCNN, which was optimized by the quantum-behaved particle swarm optimization (QPSO) algorithm. They used the PCNN model to find optimal parameters about source images for fusion. In the NSST domain, Ganasala and Kumar [38] motivated PCNN to process low-frequency (LF) and high-frequency (HF) subbands by normalized coefficient value. The generated fused images performed more details and better contrast.

III. Methodology

In this section, we describe the flexible mask-refined multimodal medical image fusion architecture in detail. At first, we introduce the overall framework of the proposed GeSeNet in Section III-A. Then, three devised modules, that is, the edge enhancement module, the region mask module, and the global refinement module, are explained in Sections III-B–III-D, respectively. Moreover, we discuss the specific representation of the loss function in Section III-E, including the edge and semantic loss functions.

A. Overall Framework

The overall framework of our proposed method is shown in Fig. 2, which consists of a multimodal medical image fusion network and a semantic module. As different inputs, MRI images combined with CT, PET, and SPECT images are employed to conduct typical multimodal medical image fusion. The size of MRI and CT images are $H \times W \times 1$, where H and W mean height and width, respectively, and 1 represents the number of channels contained in images. Similarly, $H \times W \times 3$ is the size of PET and SPECT images.

In the fusion network of Fig. 3, we first feed source images into the edge enhancement module to highlight edge textures of fusion results. The structure of this module can retain the

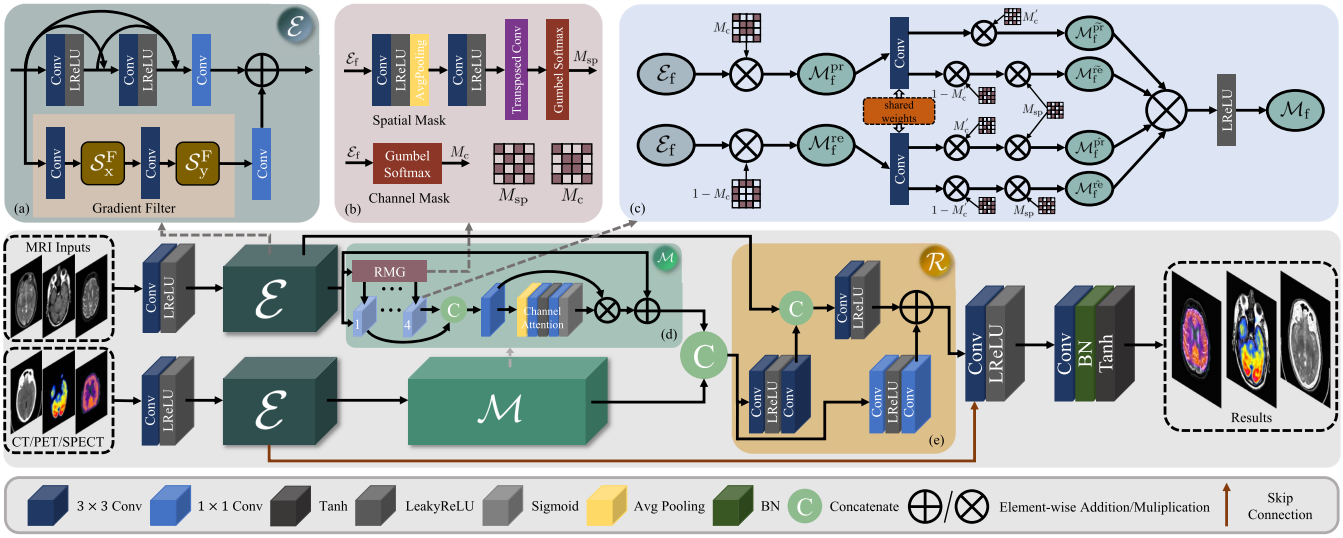

Fig. 3. Architecture of GeSeNet about our proposed method. (a) Pipeline of the edge enhancement module. (b) and (c) Detailed structure of RMG and RMC in the training phase, respectively. (d) Pipeline of the region mask module. (e) Pipeline of the global refinement module. The bottom position shows the legend of the proposed fusion network.

edge information of structural medical images to the greatest extent. The extraction process can be formulated as follows:

$$\mathcal{E}_f = \mathcal{E}_{\text{Conv}} + \mathcal{E}_G \tag{1}$$

where $\mathcal{E}_f$, $\mathcal{E}_{\text{Conv}}$, and $\mathcal{E}_G$ means extracted features, the convolutional structure, and gradient filter used in the edge enhancement module, respectively. After modification of the edge enhancement module, the extracted features are then input into the region mask module. We obtain marked "principal" and "redundant" regions from this module, which can reduce redundant computation and emphasize the representation of important features in the fusion process. The optimized features $\mathcal{M}_f$ by the region mask module from different branches are concatenated, which is calculated as follows:

$$\mathcal{M}_f = \text{Concatenate}(\mathcal{M}_a, \mathcal{M}_b) \tag{2}$$

where a and b represent the MRI branch and the CT/PET/SPECT branch, respectively. In addition, the global refinement module is initiated to revise features from different modules. We can define this process as follows:

$$\mathcal{R}_f = \mathcal{R}_{\text{Conv}}(\mathcal{E}_a, \mathcal{M}_f) \tag{3}$$

where $\mathcal{R}_f$ and $\mathcal{R}_{\text{Conv}}$ indicate the refined features and the used convolutional layers in the global refinement module, respectively.

After completing the fusion process, the initial fused images are fed into a per-trained semantic module [39] to learn semantic information. The semantic module optimizes high-level and low-level feature maps simultaneously to capture more accurate semantic information S_f from fusion results. The extraction process can be quantified as follows:

$$S_f = S(f_h, f_l) \tag{4}$$

where f_h and f_l, respectively, denote high-level and low-level feature maps. Guided by a newly designed semantic loss, the semantic module input the learned semantic information into the former fusion network through back-propagation. Due to the combination of semantic and edge-enhanced information, the edge details of fusion results can be highlighted more obviously.

When fusing RGB three-channel source images, that is, PET and SPECT images, we convert them to YCbCr three-channel form for fusion. Specifically, we first fuse the luminance information in the Y channel with a single-channel MRI image to generate a gray-scale fusion result. Owing to the content features and details in the Y channel, the vital information from source images can be retained substantially on fused images. Then, the chrominance information on the Cb and Cr channels are combined through a quantitative formula, which the result is used as the color representation of the fused image

$$C_f = \frac{Cb_i(|Cb_i - \tau|) + Cr_i(|Cr_i - \tau|)}{|Cb_i - \tau| + |Cr_i - \tau|} \tag{5}$$

where C_f means the weighted sum result in fusion images. Cb_i and Cr_i are the chrominance values of each pixel in source images. Inspired by previous works [40], [41], we also set the hyperparameter τ to 128. At last, we fuse the result in the Y channel with C_f to obtain a YCbCr 3-channel result and convert it into the RGB form.

B. Edge Enhancement Module

In multimodal medical image fusion, significant edge information can make it easier for researchers and doctors to conduct scientific research and pathological analysis [18]. In this case, we propose an edge enhancement module to strengthen the edge representation of fused images, which makes both qualitative performance and quantitative metrics achieve a higher level.

As shown in Fig. 3(a), we were inspired by the structure of the resblock [42] to design the edge enhancement module. In the mainstream, two 3×3 convolutional layers with dense connection mode and a 1×1 convolutional layer

are employed to extract shallow feature maps $\mathcal{E}_{\text{Conv}}$ from source images. We use the leaky rectified linear unit (LReLU) as their activation function. In the residual stream, a novel gradient filter is introduced for learning gradient information. The gradient filter first deploys a 3×3 convolutional layer with a Sobel operator in the horizontal direction to calculate horizontal gradient magnitude. Similarly, we can get vertical gradient magnitude from another 3×3 convolutional layer and a vertical Sobel operator. Then, the gradient information $\mathcal{E}_G$ is fed into a 1×1 convolutional layer to remove differences of channel dimensional. Furthermore, we implement element-wise addition to merge $\mathcal{E}_{\text{Conv}}$ and $\mathcal{E}_G$ to obtain the final edge-enhanced features $\mathcal{E}_f$.

In addition, to prevent the proposed GeSeNet from forgetting learned edge information during training, we conduct a skip connection operation on the CT/PET/SPECT branch to connect with the latter convolutional layer. The edge-enhanced features $\mathcal{E}_f$ on the MRI branch are exploited in the global refinement module to revise some details of fusion images, which are given a detailed explanation in Section III-D.

C. Region Mask Module

Mask techniques are well used in many computer vision tasks, for example, image super-resolution [43], target detection [44], and semantic segmentation [45]. To complete the multimodal medical image fusion task more efficiently, a region mask module is introduced in GeSeNet.

After integrating edge gradient features, the region mask module in Fig. 3(d) is arranged to divide two different regions for avoiding redundant computation and extracting fine-grained details from source images. Due to the different purposes between the training and testing phases, we manipulate corresponding structures of the region mask module that contains a region mask generator (RMG) and four region mask convolutions (RMCs) to learn and optimize features, respectively.

1) Region Mask Generator: In the training phase, the spatial and channel masks are employed to mark the "principal" and "redundant" regions in RMG, respectively. As shown in Fig. 3(b), the edge-enhanced features $\mathcal{E}_f$ are first fed into a 3×3 convolutional layer with LReLU and average-pooling layers. After modification by another 3×3 convolutional layer with LReLU, we input the extracted feature maps into a transposed convolution to obtain the upsampled feature $\mathcal{E}_f^{\text{sp}}$. The learnable parameter from the transposed convolution can be updated through back-propagation so that the upsampling operation can be performed efficiently. To realize the spatial mask self-regulating, Gumbel softmax distribution is introduced to estimate a one-hot distribution [46], which can be formulated as follows:

$$M_{\text{sp}} = \frac{\exp\left((\mathcal{E}_f^{\text{sp}}[1 \times h \times w] + G_{\text{sp}}[1 \times h \times w])/\theta\right)}{\sum_{i=1}^{2} \exp\left((\mathcal{E}_f^{\text{sp}}[i \times h \times w] + G_{\text{sp}}[i \times h \times w])/\theta\right)} \tag{6}$$

where h and w represent factors in the vertical and horizontal directions, respectively. G_{sp} is an intermediate noise tensor in Gumbel softmax, in which all elements obey Gumbel distribution [$\mathbb{R} \in (0, 1)$]. When θ tends to ∞, feature maps may perform uniform distribution. Furthermore, when θ tends to 0, one-hot distribution may appear. Owing to the constraint of the range, θ as a hyperparameter is set to 0.5 for balance. To mark "redundant" regions by the channel mask M_c, we randomly initialize $\mathcal{E}_f$ to $\mathcal{E}_f^c$ on Gaussian distribution [$\mathbb{R} \in (0, 1)$] before feeding into Gumbel softmax. M_c is defined as follows:

$$M_c = \frac{\exp\left((\mathcal{E}_f^c[1 \times c] + G_c[1 \times c])/\theta\right)}{\sum_{i=1}^{2} \exp\left((\mathcal{E}_f^c[i \times c] + G_c[i \times c])/\theta\right)} \tag{7}$$

where c means the number of channels.

During the testing phase, an Argmax layer is introduced to substitute the Gumbel softmax to obtain spatial mask and channel mask in Fig. 4(a). The Argmax layer can return the corresponding index value of the maximum value in features.

2) Region Mask Convolution: To obtain the mask-optimized feature $\mathcal{M}_f$, we input the extracted spatial mask and channel mask into RMC as shown in Fig. 3(c). Specifically, we introduce four RMC structures with dense connections. First, $\mathcal{E}_f$ separately performs element-wise multiplication with M_c and $1 - M_c$ to get initial "principal" and "redundant" features, that is, $\mathcal{M}_f^{\text{pr}}$ and $\mathcal{M}_f^{\text{re}}$. In other words, we divide these features into two distinct regions. Then, $\mathcal{M}_f^{\text{pr}}$ and $\mathcal{M}_f^{\text{re}}$ are fed into two 3×3 convolutional layers with shared weights to implement element-wise multiplication with M_c', $1 - M_c'$, and M_{sp}. At last, we integrate all the features from different regions to generate the mask-optimized feature $\mathcal{M}_f$. Moreover, the gradient of all features can be retained by Gumbel softmax to adjust the kernel weights in RMC.

In the testing phase, we introduce sparse convolutional layers to deal with the spatial and channel masks in Fig. 4(b) and (c). Through M_c and M_c', the kernel of RMC can get four split subkernels, that is, one 3×3 convolution and three 3×3 sparse convolutions. With the action of M_c', $\mathcal{M}_f^{\text{pr}}$ and $\mathcal{M}_f^{\text{re}}$ can be achieved. We employ the 3×3 convolution and the sparse convolution 1 with M_{sp} to get $\mathcal{M}_f^{\tilde{\text{pr}}}$ and $\mathcal{M}_f^{\tilde{\text{re}}}$. Meanwhile, $\mathcal{M}_f^{\hat{pr}}$ and $\mathcal{M}_f^{\hat{re}}$ can be obtained as the similar approach with sparse convolutions 2 and 3. Finally, we use element-wise multiplication and concatenate operation to generate $\mathcal{M}_f$.

As shown in Fig. 3(d), a channel attention module is exploited to further modify the mask-optimized feature $\mathcal{M}_f$. The selection of activate functions is LReLU and Sigmoid. In the proposed fusion network, we concat $\mathcal{M}_f$ from different branches before the global refinement module.

D. Global Refinement Module

In the global refinement module, we first import $\mathcal{M}_f$ into two 3×3 convolutions to get global mask-optimized features $\mathcal{M}_f^g$. Then, the edge-enhanced features $\mathcal{E}_f$ from the MRI branch are concatenated with $\mathcal{M}_f^g$ for reserving the former extracted information to refine the performance of fusion results. Finally, we integrate all features to generate the global-refined feature by utilizing element-wise addition, which uses two 1×1 convolutional layers to remove channel

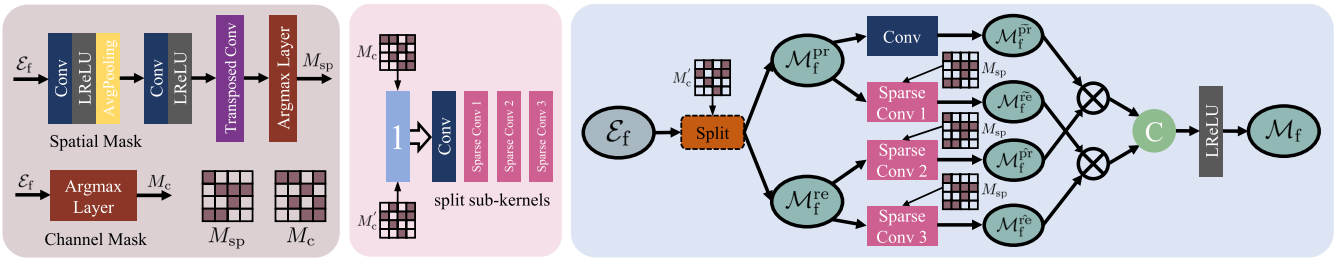

Fig. 4. Structure of the region mask module in the testing phase: (a) pipeline of RMG, (b) and (c) pipeline of RMC with sparse convolution. Instead of using the shared weight strategy in the training phase, the split manipulation is deployed in RMC during the testing phase.

dimensional diversities. The refinement process can be quantified as follows:

$$\mathcal{R}\left(\mathcal{E}_f|\mathcal{M}_f^g, \mathcal{M}_f\right) = \mathcal{E}_f \odot \mathcal{M}_f^g + \mathcal{M}_f \tag{8}$$

where $\odot$ indicates the concatenate operation. In addition, to make fusion results vivid in color and retain more functional details, a skip connection is introduced to connect the edge enhancement module in the CT/PET/SPECT branch with the 3×3 convolution behind the global refinement module.

E. Loss Function

To ensure the quality of fusion images by persisting more meaningful extracted information, the loss function of our proposed method consists of three parts, which contain the edge loss $\mathcal{L}_E$, the structure similarity index measure loss $\mathcal{L}_{\text{SSIM}}$ and the semantic loss $\mathcal{L}_S$. The total loss $\mathcal{L}_{\text{total}}$ is defined as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_E + \alpha\mathcal{L}_{\text{SSIM}} + \beta\mathcal{L}_S \tag{9}$$

where α and β are tradeoff hyperparameters to balance the values of $\mathcal{L}_{\text{total}}$.

1) Edge Loss: During the training phase, the edge loss can guide source images to generate a fused image that highlights content performance and gradient information. Therefore, we divide the edge loss into two parts, which can be described as follows:

$$\mathcal{L}_E = \mathcal{L}_c + \gamma\mathcal{L}_g \tag{10}$$

where $\mathcal{L}_c$ and $\mathcal{L}_g$ denote the content and gradient loss, respectively. γ is a hyperparameter for controlling the value of $\mathcal{L}_g$.

In the content loss, we employ the l_1-norm to measure the difference between generated outputs I_F and source images. $\mathcal{L}_c$ is formulated as follows:

$$\mathcal{L}_c = \frac{1}{H_F \cdot W_F} \cdot \|I_F - \max(I_a, I_b)\|_1 \tag{11}$$

where H_F and W_F represent the height and width of I_F, respectively. $\max(*)$ and $\|*\|_1$ mean the maximum selection strategy and l_1-norm, respectively. According to the content loss, the pixel-level contact information is transferred to our proposed network for image fusion.

We expect to retain more edge textures while delivering the content information. Hence, the gradient loss is proposed to measure the value of the gradient in the pixel domain, which can be calculated as follows:

$$\mathcal{L}_g = \frac{1}{H_F \cdot W_F} \cdot \left\| |\nabla I_F| - \max(|\nabla I_a|, |\nabla I_b|) \right\|_1 \tag{12}$$

where ∇ denotes the Sobel operator to calculate the value of the gradient. Since negative gradients are not available, absolute value operation $|*|$ is introduced to solve this problem.

2) Structure Similarity Index Measure Loss: $\mathcal{L}_{\text{SSIM}}$ can measure the structural difference by structural similarity index measure (SSIM) [47], which contains three kinds of information, that is, luminance, structure, and contrast. We can specifically express $\mathcal{L}_{\text{SSIM}}$ as follows:

$$\mathcal{L}_{\text{SSIM}} = (1 - \text{SSIM}(I_F, I_a)) + (1 - \text{SSIM}(I_F, I_b)). \tag{13}$$

Moreover, $\text{SSIM}(I_F, I_*)$ is defined as follows:

$$\sum_{I_F, I_*} \frac{2\mu_{I_F}\mu_{I_*} + C_1}{\mu_{I_F}^2 + \mu_{I_*}^2 + C_1} \cdot \frac{2\sigma_{I_F}\sigma_{I_*} + C_2}{\sigma_{I_F}^2 + \sigma_{I_*}^2 + C_2} \cdot \frac{\sigma_{I_F I_*} + C_3}{\sigma_{I_F}\sigma_{I_*} + C_3} \tag{14}$$

where I_* indicates the source image I_a or I_b. μ and σ mean the average value and standard deviation (SD), respectively. C_1, C_2, and C_3 are constants for steadying the indicator.

3) Semantic Loss: The semantic loss is introduced to feed semantic information from source images into fusion results. Inspired by the previous work [39], we separate the semantic loss into the main semantic loss $\mathcal{L}_{\text{main}}$ and the subsidiary semantic loss $\mathcal{L}_{\text{sub}}$, which can be shown as follows:

$$\mathcal{L}_S = \mathcal{L}_{\text{main}} + \delta\mathcal{L}_{\text{sub}} \tag{15}$$

where δ can keep the value of $\mathcal{L}_S$ stable. Furthermore, the main semantic loss and the subsidiary semantic loss can be, respectively, defined as follows:

$$\mathcal{L}_{\text{main}} = -\frac{1}{H_F \cdot W_F} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} (I_o \cdot \ln S_{\text{main}}) \tag{16}$$

$$\mathcal{L}_{\text{sub}} = -\frac{1}{H_F \cdot W_F} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C} (I_o \cdot \ln S_{\text{sub}}) \tag{17}$$

where I_o denotes a one-hot distribution generated by extracted semantic features. S_{main} and S_{sub} mean the main semantic and the subsidiary semantic information, respectively.

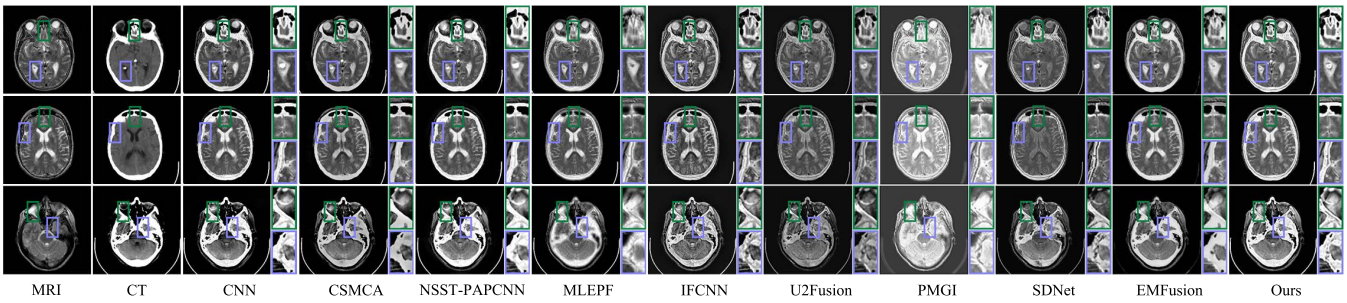

Fig. 5. Qualitative comparison results of our GeSeNet with nine state-of-the-art methods on three MRI and CT image pairs. MLEPF [48] is introduced to replace DDcGAN [49] for MRI-CT fusion. Two enlarged detail patches are shown on the right of each result. Noticeably, our method can obtain well-performance results that contain abundant structural information.

IV. Experiment

In this section, we first give the specific experimental details, qualitative comparison approaches, and quantitative evaluation metrics. Second, MRI-CT, MRI-PET, and MRI-SPECT comparison results are shown to demonstrate the superiority of GeSeNet. In addition, we compare the time efficiency and parameter quantity of each method. Finally, ablation experiments are conducted to validate the effectiveness of the devised modules and loss functions.

A. Experimental Details

The training and test datasets of our proposed method are selected on the *Harvard* medical dataset, which can be publicly available at http://www.med.harvard.edu/AANLIB/home.html. Specifically, we choose 150 image pairs from MRI-CT, MRI-PET, and MRI-SPECT images and crop them into patches with size 24×24 to treat as the training dataset. Twenty-one pairs of MRI-CT images, 42 pairs of MRI-PET images, and 73 pairs of MRI-SPECT images which can typically highlight different characteristics are regarded as test datasets to complete different medical image fusion tasks. Note that the experimental setup of three different modalities keeps uniform in the training and testing phases.

During training, the Adam optimizer is employed with the stride of 8, the batch size of 4, the original learning rate of 1e-3, and the weight decay of 2e-4 to train the proposed fusion network. In the semantic module, we use stochastic gradient descent with a batch size of 4, a momentum of 0.9, and a weight decay of 5e-4 to obtain semantic information after training the fusion processing. The epoch is set to 500. The used convolutions and activate functions in the proposed method are performed in the legend of Fig. 3. To calculate the value of the loss function easily, we preset hyperparameters γ and δ to 10 and 0.5, respectively. According to previous works [50], [51], α is set to 0.5. Furthermore, we set β to 0.3 for balancing the extracted information between the fusion network and the semantic module. The detailed experiment is described in Section IV-G. All experiments are deployed in the PyTorch framework with a PC, which has an NVIDIA GeForce RTX 3060 GPU, a 16-GB RAM memory, and an Intel Core i5-11400F CPU.

B. Comparison Approaches and Evaluation Metrics

1) Comparison Approaches: We compare our proposed methods with ten state-of-the-art methods, which contain one traditional method, that is, CSMCA [16], two PCNN-based methods, that is, NSST-PAPCNN [1] and MLEPF [48], and seven deep-learning-based methods, that is, CNN [32], DDcGAN [49], IFCNN [34], U2Fusion [12], PMGI [52], SDNet [53], and EMFusion [13]. It is worth noting that DDcGAN is proposed to fuse structural images with functional images so that we introduce MLEPF for MRI-CT fusion instead. Moreover, the comparison approaches are all publicly available and we set the same parameters as the original papers during the testing phase.

2) Evaluation Metrics: To quantify the merits of our fusion results, we select six quantitative evaluation metrics, that is, SSIM [47], SD, mutual information (MI) [54], visual information fidelity (VIF) [55], the sum of the correlations of differences (SCDs) [56], and edge-based similarity measure ($Q_{ab/f}$) [57] to compare with other ten state-of-the-art methods. Specifically, SSIM is a unified metric that is used to measure the similarity of two images. The measured information is luminance, structure, and contrast. In other words, the higher the value of SSIM, the more similar the two images are. SD can represent the degree of offset between the pixel value and the average pixel value of the image. From a statistical view, SD means the distribution and contrast of images. Based on the knowledge of information theory, MI can calculate the amount of information interaction from source images to the fused image. Generally, a larger value of MI symbolizes that the fused result has a better performance. VIF quantifies the information fidelity of visual perception, which is consistent with the human visual system. By building a complex model, we calculate the degree of distortion between the source images and the fused image to get the VIF. A large VIF indicates a high level of information fidelity. SCD targets to measure the differences between source images and the fusion image in each pixel. A high score in SCD represents an excellent fusion result. $Q_{ab/f}$ aims to calculate the total edge information which is transferred from source images to the output.

C. MRI-CT Comparison Results

1) Qualitative Analysis: As shown in Fig. 5, we perform three typical sets of MRI-CT fusion results to compare our

TABLE I

QUANTITATIVE COMPARISON OF OUR GESENET WITH OTHER NINE METHODS ON THE MRI-CT TEST DATASET. THE BEST AVERAGE VALUE AND SD ARE MARKED IN RED WITH BOLD FONT AND THE SECOND ONES ARE BOLDED IN BLUE, RESPECTIVELY

Method	$\mathbf{SSIM}_{\uparrow}$	$\mathbf{SD}_{\uparrow}$	$\mathbf{MI}_{\uparrow}$	$\mathbf{VIF}_{\uparrow}$	$\mathbf{SCD}_{\uparrow}$	$\mathbf{Q_{ab/f\uparrow}}$
CNN	0.940 ± 0.007	9.665 ± 0.430	3.127 ± 0.157	**0.744 ± 0.034**	1.286 ± 0.122	**0.597 ± 0.034**
CSMCA	0.919 ± 0.016	9.761 ± 0.361	3.010 ± 0.184	0.699 ± 0.024	1.084 ± 0.153	0.579 ± 0.040
NSST-PAPCNN	**0.949 ± 0.008**	9.837 ± 0.304	3.207 ± 0.168	0.697 ± 0.052	**1.453 ± 0.072**	0.573 ± 0.039
MLEPF	0.865 ± 0.037	**10.177 ± 0.311**	3.187 ± 0.190	0.627 ± 0.041	1.263 ± 0.092	0.454 ± 0.036
IFCNN	0.927 ± 0.009	9.363 ± 0.193	3.050 ± 0.128	0.569 ± 0.029	1.048 ± 0.133	0.586 ± 0.034
U2Fusion	0.920 ± 0.009	9.667 ± 0.432	3.029 ± 0.146	0.471 ± 0.037	0.750 ± 0.085	0.457 ± 0.014
PMGI	0.794 ± 0.046	10.117 ± 0.296	3.007 ± 0.189	0.506 ± 0.031	1.351 ± 0.081	0.416 ± 0.061
SDNet	0.909 ± 0.010	9.906 ± 0.367	3.112 ± 0.129	0.490 ± 0.026	0.970 ± 0.041	0.473 ± 0.014
EMFusion	0.899 ± 0.025	10.060 ± 0.342	**3.410 ± 0.210**	0.718 ± 0.053	1.317 ± 0.105	0.495 ± 0.025
Ours	**0.943 ± 0.008**	**10.184 ± 0.359**	**3.365 ± 0.132**	**0.720 ± 0.028**	**1.426 ± 0.088**	**0.676 ± 0.021**

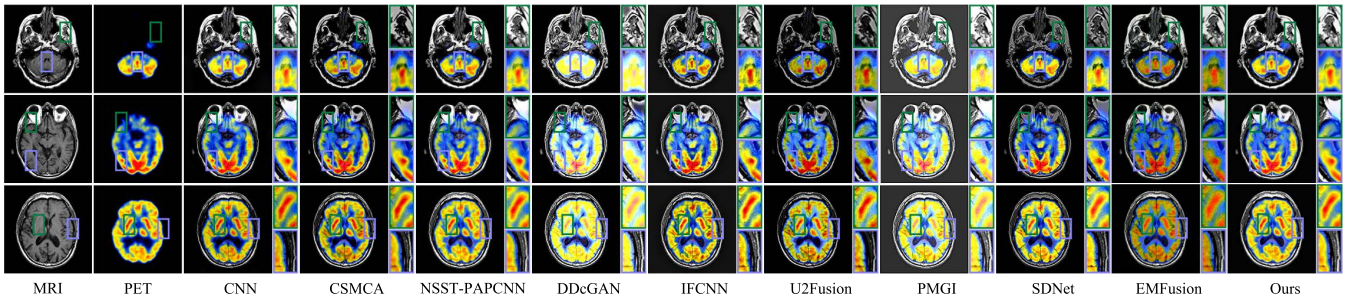

Fig. 6. Qualitative comparison results of our GeSeNet with nine state-of-the-art methods on three MRI and PET image pairs. The two magnified details are marked by green and purple boxes and shown to the right of each fusion result. Clearly, our fusion results are superior in both local details and global effects, for example, the blood flow information in the third row.

qualitative performance with the other nine state-of-the-art methods. Due to the limitations of fusion strategies, the skeleton information is unable to be visibly exhibited in the results of CSMCA, U2Fusion, and SDNet. IFCNN effectively extracts features of MRI and CT source images and generates a fusion image with rich texture details. However, the luminance of the results looks dimmer than our fusion results. PMGI cannot balance the contrast of fusion results, which is difficult for researchers and physicians to distinguish the different information represented in the fusion results. The structure of MLEPF is insensitive to edge features and detail information so that the generated images occur some distortions. In the third row, CNNs can perform visually vivid fused images, but the expressed synaptic information is slightly inferior compared to our fusion results. Though the overall visual effect and information retention of CNN and EMFusion are multiply shown in their fusion images, there also appear some artifact halos. Moreover, when CT images are less informative (e.g., in the second row), our network can retain more features from MRI images to optimize the quality of fusion results. We attribute this advantage to the region mask module.

2) Quantitative Analysis: For a more comprehensive comparison, we give the six aforementioned evaluation indicators in Table I. Our fusion results achieve the best marks on SD and $Q_{ab/f}$, and the second best on SSIM, MI, VIF, and SCD. From the quantitative indexes, the higher SD value indicates that our generated results can stay more stable existence at the pixel level. Owing to devising the edge enhancement module, fused results are sensitive to the performance of edge details, and the value of $Q_{ab/f}$ can get the best grade. In the MRI-CT test dataset, structural information contained in some CT images (e.g., the first and second rows in Fig. 5) is not obvious to capture. Unlike other methods, our proposed GeSeNet introduces the region mask module to mark "principle" regions and extract more features from MRI images for guaranteeing the quality of the fused results. This is why our method is suboptimal on some metrics. As a result, the proposed method keeps the quantitative indicators in an excellent position while ensuring the visual effect of the fused images.

D. MRI-PET Comparison Results

1) Qualitative Analysis: We conduct a subjective qualitative comparison of MRI-PET fusion in Fig. 6. It is evident that our fusion results not only highlight different characteristics of MRI and PET source images, but also show faithful color representation. DDcGAN is a method developed based on GAN, so that unstable blurring artifacts may appear in the fusion results. As a unified fusion framework, PMGI may occur an unbalanced weight distribution ratio during fusing medical images, which leads to the vital information in source images cannot be completely transmitted to the corresponding generated results. The results of U2Fusion perform weaker information extraction and color realization compared with GeSeNet. CSMCA and SDNet are better than U2Fusion in color representation, however, the texture details on MRI images are still preserved poorly in their fusion results. Though EMFusion can extract more MRI information, the performance of color is distorted in fused results. For other comparison methods, GeSeNet outperforms in both edge preservation and texture rendering, which can emphasize soft-tissue structure and functional information simultaneously.

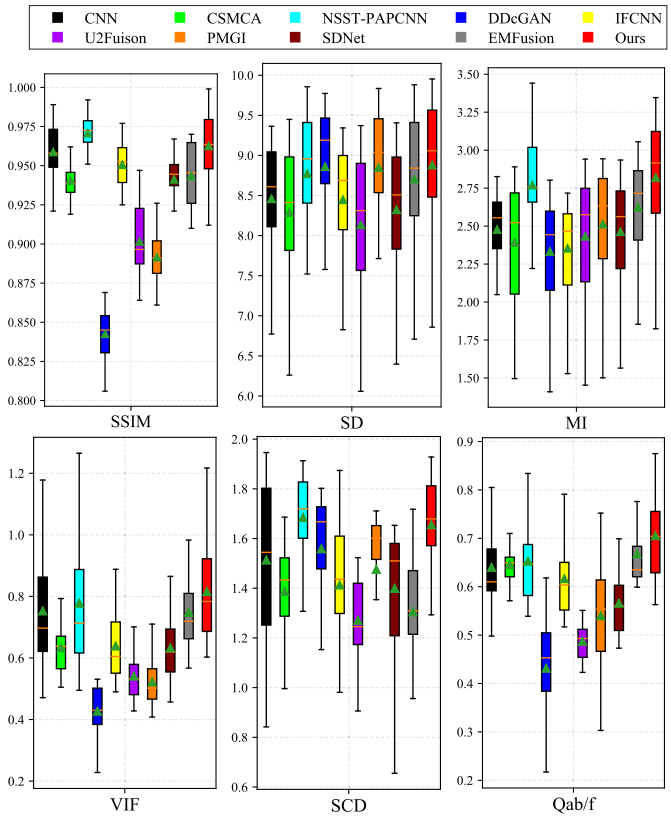

Fig. 7. Quantitative comparison of our GeSeNet with other nine methods on the MRI-PET test dataset. The green triangles in each rectangle represent the mean value of different methods.

2) Quantitative Analysis: Fig. 7 shows mean value (represented by green triangles in each rectangle), SD (represented by rectangle length), median number (represented by orange lines in each rectangle), and fluctuation range (represented by the total length of the line) of six evaluation metrics on the MRI-PET test dataset. From the statistical results, it can be seen that the results generated by our method achieve the largest averages on SD, MI, VIF, and $Q_{ab/f}$, which denotes that the proposed method transfers more useful information from source images and performs more abundant texture details to researchers. For the metrics SSIM and SCD, the average value of our fusion results obtains the second best score. Specifically, the SSIM and SCD mean values not reaching the highest level does not mean that our fusion results are of poor quality. Due to the labeling of different regions and the targeted extraction of features, the generated fusion images may miss some minor information and reduce the performance of some quantitative indicators.

E. MRI-SPECT Comparison Results

1) Qualitative Analysis: Quantitative comparison results about the MRI-SPECT fusion task are shown in Fig. 8. Similar to MRI-PET fusion, our method also exhibits vivid colors and rich texture details on the MRI-SPECT task. In the second row, the proposed network transfers the structural information from MRI well into the fused images, while also attaching the functional information from SPECT images to the fusion results. It can ensure that detailed features are not covered by the chrominance information. At the green patch in the third row, we can clearly observe that when there are two kinds of different information at the same position, our method can realize that the two kinds of information exist on a fused image at the same time.

2) Quantitative Analysis: As shown in Fig. 9, we present quantitative comparison results in the form of scatter plots. The horizontal and vertical coordinates in Fig. 9 separately represent the six evaluation indicators. Since the selected indicators are all positively correlated, the farther the marked point is from the axis, the better its performance. Apparently, the value of SD, SSIM, MI, SCD, and $Q_{ab/f}$ achieve the highest score compared with other methods, which indicates that the proposed GeSeNet has outstanding performance on similarity preservation and information transfer. The value of VIF is slightly lower than CNNs and gets a suboptimal score. However, it does not affect the quality of our fusion result. In color performance, our proposed method outperforms CNNs.

F. Efficiency Comparison

In addition to comparing qualitative and quantitative results generated by the models, the time efficiency and size of the models are also critical indicators to evaluate the quality of the proposed method. As shown in Table II, we perform the average runtime of three tasks (i.e., MRI-CT, MRI-PET, and MRI-SPECT) and parameter quantity of the above-mentioned comparison methods. The traditional and PCNN-based methods are all operated with MATLAB on an i5-11400F CPU. Besides CNN running on CPU, other deep-learning-based methods are performed with Tensorflow/PyTorch on an NVIDIA GeForce RTX 3060 GPU.

In terms of runtime, our proposed method achieves the shortest time on the test dataset. Due to the end-to-end structure, the proposed method can reduce the tediousness of manually regulating the fusion strategy. Moreover, we design the corresponding architecture and network of our proposed method to extract and fuse more efficiently through the essence of multimodal medical images, which can avoid running inefficiencies caused by directly adding convolutional layers. On hardware systems, methods running on GPUs tend to run more efficiently than CPUs. The parameter size of the proposed method stands intermediate level. Owing to the framework of the region mask module, the "principal" and "redundant" regions are divided to focus on learning important information while ignoring redundant information. Though the space complexity of our proposed method is not the best, the time efficiency and the fusion quality perform excellently above these mentioned methods.

G. Ablation Experiments

1) Analysis on Different Modules: We analyze the model architecture of the proposed network and sequentially verify the effectiveness of each module in our approach. To simplify the analysis process, the whole network is divided into three main parts, that is, $\mathcal{E}$, $\mathcal{M}$, and $\mathcal{R}$, which means the

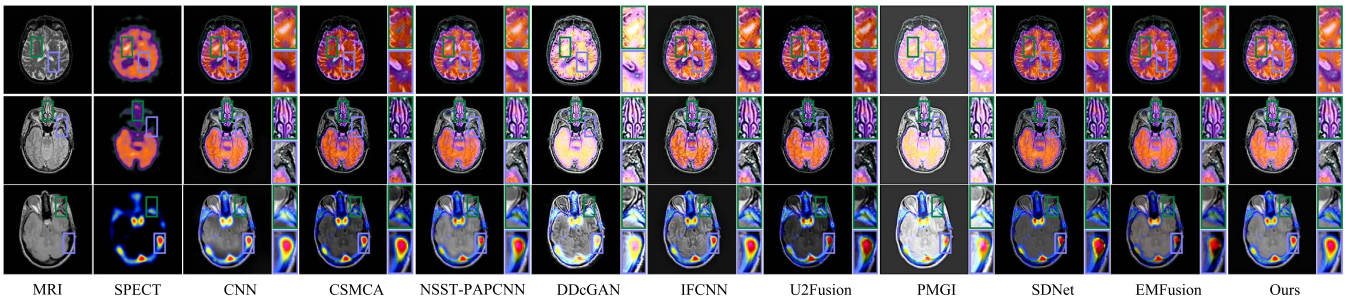

Fig. 8. Qualitative comparison results of our GeSeNet with nine state-of-the-art methods on three MRI and SPECT image pairs. The two magnified details are marked by green and purple boxes and shown to the right of each fusion result. We can clearly observe that the proposed method performs more vividly than other compared methods, for example, the bone junction in the third row.

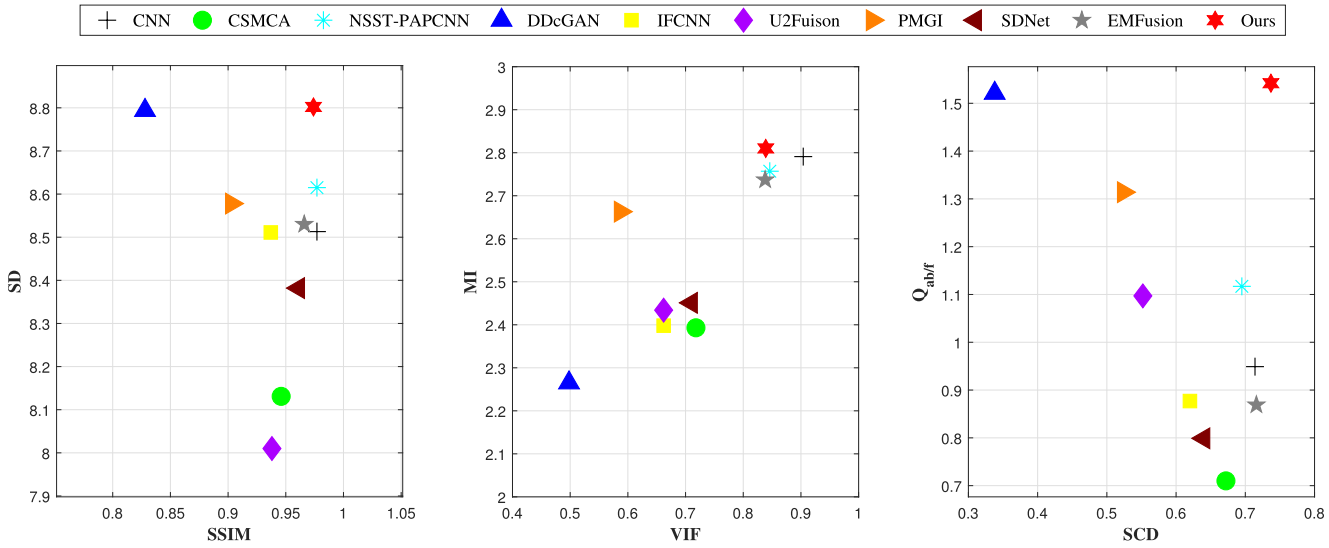

Fig. 9. Quantitative comparison of our proposed method with other nine methods on MRI-SPECT test dataset. Our method is in a leading position.

TABLE II

AVERAGE RUNTIME (UNIT: SECOND) AND PARAMETER SIZE (UNIT: MB) OF DIFFERENT METHODS. CNN, TRADITIONAL, AND PCNN-BASED METHODS ARE PERFORMED ON CPU, WHILE OTHER METHODS ARE CONDUCTED ON GPU

Method	CSMCA	NSST-PAPCNN	MLEPF	CNN	DDcGAN	IFCNN	U2Fusion	PMGI	SDNet	EMFusion	Ours
Device	CPU	CPU	CPU	CPU	GPU	GPU	GPU	GPU	GPU	GPU	GPU
Runtime$_{\downarrow}$	60.276	3.294	9.178	9.386	0.873	**0.035**	0.264	0.074	0.056	0.172	**0.011**
Parameters$_{\downarrow}$	-	-	-	0.500	1.908	0.114	0.659	**0.042**	**0.067**	0.297	0.240

edge enhancement module, the region mask module, and the global refinement module, respectively. As shown in Fig. 10, we present qualitative results of the proposed network with or without each mentioned module. It is worth noting that the introduction of any module in the network has a good effect on the quality of the generated results. In detail, the fused results may appear blur edge and detail absence without $\mathcal{E}$. In the MRI-CT fusion results, it is not difficult to observe that the full model can obtain more edge details than the model without $\mathcal{E}$. Hence, it shows that $\mathcal{E}$ is sensitive to edge information in the network. $\mathcal{M}$ can distinguish valid and invalid information for more targeted capture features. Some useless information (e.g., patch enlarged in green box of the MRI-SPECT result) may occur in the fused images to affect the overall visual effect without $\mathcal{M}$. The full model divides different regions with the help of $\mathcal{M}$ to focus on extracting important features while reducing the reuse of redundant information. $\mathcal{R}$ plays a role in modifying the overall performance of results. As shown in the MRI-PET result, it fails to reserve texture details and global perception without $\mathcal{R}$. In addition to qualitative analysis, we conduct three sets of quantitative comparisons about with or without each module in Table III. By integrating different modules into our method, the full model leads the way in quantitative metrics. As a result, each module contributes positively to the final performance of the fused image. Moreover, Fig. 11 shows the visual illustrations of the region mask module, which can prove that it can accurately mark "principal" regions in different features.

2) Analysis on Module Location: The visual effect of the fused images is also related to the location of the employed modules. As the global refinement module targets to integrate the former extracted features, we only exchange the

TABLE III

ABLATION QUANTITATIVE EXPERIMENT OF EACH MODULE. THE OPTIMAL AND SUBOPTIMAL RESULTS ARE BOLDED AND MARKED IN RED AND BLUE, RESPECTIVELY

Fusion task		$\mathcal{E}$	$\mathcal{M}$	$\mathcal{R}$	Metric					
					SSIM$_\uparrow$	SD$_\uparrow$	MI$_\uparrow$	VIF$_\uparrow$	SCD$_\uparrow$	Q$_{ab/f\uparrow}$
MRI and CT	M1	✗	✔	✔	0.885 ± 0.014	9.974 ± 0.334	**3.287 ± 0.145**	**0.714 ± 0.039**	1.374 ± 0.103	0.587 ± 0.040
	M2	✔	✗	✔	**0.940 ± 0.004**	**10.028 ± 0.385**	3.231 ± 0.141	0.661 ± 0.034	**1.418 ± 0.085**	**0.612 ± 0.036**
	M3	✔	✔	✗	0.920 ± 0.006	9.944 ± 0.403	3.171 ± 0.156	0.645 ± 0.035	1.395 ± 0.051	0.561 ± 0.040
	M4	✔	✔	✔	**0.943 ± 0.008**	**10.184 ± 0.359**	**3.365 ± 0.132**	**0.720 ± 0.028**	**1.426 ± 0.088**	**0.676 ± 0.021**
MRI and PET	M1	✗	✔	✔	0.904 ± 0.025	8.674 ± 0.896	**2.680 ± 0.405**	**0.826 ± 0.151**	**1.621 ± 0.184**	**0.655 ± 0.105**
	M2	✔	✗	✔	**0.941 ± 0.019**	**8.681 ± 0.893**	2.514 ± 0.359	0.734 ± 0.147	1.611 ± 0.182	0.611 ± 0.067
	M3	✔	✔	✗	0.909 ± 0.019	8.404 ± 0.924	2.438 ± 0.362	0.705 ± 0.128	1.615 ± 0.152	0.562 ± 0.079
	M4	✔	✔	✔	**0.963 ± 0.020**	**8.882 ± 0.888**	**2.821 ± 0.405**	**0.816 ± 0.168**	**1.656 ± 0.216**	**0.706 ± 0.086**
MRI and SPECT	M1	✗	✔	✔	0.904 ± 0.022	**8.635 ± 0.742**	**2.624 ± 0.283**	**0.744 ± 0.095**	1.105 ± 0.268	0.632 ± 0.096
	M2	✔	✗	✔	**0.963 ± 0.008**	8.565 ± 0.798	2.597 ± 0.258	0.713 ± 0.079	**1.392 ± 0.128**	**0.637 ± 0.057**
	M3	✔	✔	✗	0.926 ± 0.015	8.310 ± 0.852	2.431 ± 0.326	0.625 ± 0.079	1.321 ± 0.119	0.576 ± 0.058
	M4	✔	✔	✔	**0.974 ± 0.008**	**8.802 ± 0.786**	**2.810 ± 0.266**	**0.839 ± 0.082**	**1.542 ± 0.100**	**0.737 ± 0.048**

TABLE IV

ABLATION QUANTITATIVE EXPERIMENT OF LOSS FUNCTIONS. WE BOLD OPTIMAL AND SUBOPTIMAL RESULTS IN RED AND BLUE, RESPECTIVELY

Metric	MRI and CT			MRI and PET			MRI and SPECT		
	w/o $\mathcal{L}_E$	w/o $\mathcal{L}_S$	Ours	w/o $\mathcal{L}_E$	w/o $\mathcal{L}_S$	Ours	w/o $\mathcal{L}_E$	w/o $\mathcal{L}_S$	Ours
SSIM$_\uparrow$	0.840 ± 0.022	**0.945 ± 0.005**	**0.943 ± 0.008**	0.803 ± 0.055	**0.942 ± 0.020**	**0.963 ± 0.020**	0.829 ± 0.033	**0.963 ± 0.010**	**0.974 ± 0.008**
SD$_\uparrow$	9.986 ± 0.380	**10.017 ± 0.384**	**10.184 ± 0.359**	8.360 ± 0.890	**8.773 ± 0.895**	**8.882 ± 0.888**	8.371 ± 0.877	**8.634 ± 0.783**	**8.802 ± 0.786**
MI$_\uparrow$	3.092 ± 0.167	**3.314 ± 0.138**	**3.365 ± 0.132**	2.302 ± 0.358	**2.591 ± 0.385**	**2.821 ± 0.405**	2.395 ± 0.278	**2.641 ± 0.252**	**2.810 ± 0.266**
VIF$_\uparrow$	0.575 ± 0.035	**0.709 ± 0.035**	**0.720 ± 0.028**	0.581 ± 0.124	**0.769 ± 0.152**	**0.816 ± 0.168**	0.574 ± 0.062	**0.746 ± 0.089**	**0.839 ± 0.082**
SCD$_\uparrow$	1.164 ± 0.153	**1.358 ± 0.105**	**1.426 ± 0.088**	0.887 ± 0.164	**1.585 ± 0.207**	**1.656 ± 0.216**	0.515 ± 0.257	**1.175 ± 0.263**	**1.542 ± 0.100**
Q$_{ab/f\uparrow}$	0.244 ± 0.022	**0.617 ± 0.033**	**0.676 ± 0.021**	0.187 ± 0.040	**0.639 ± 0.076**	**0.706 ± 0.086**	0.140 ± 0.023	**0.658 ± 0.066**	**0.737 ± 0.048**

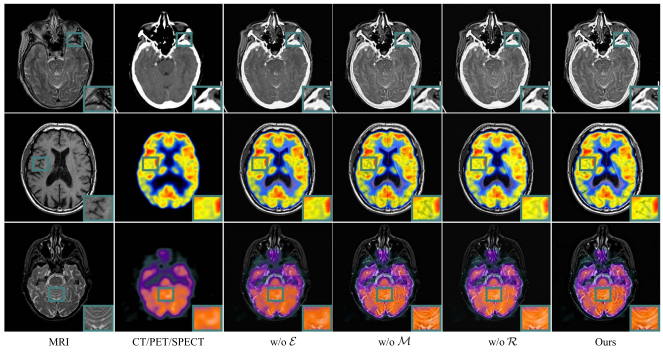

Fig. 10. Ablation qualitative experiment of each module on three kinds of image pairs, that is, MRI-CT, MRI-PET, and MRI-SPECT images. Each module plays an active role in GeSeNet.

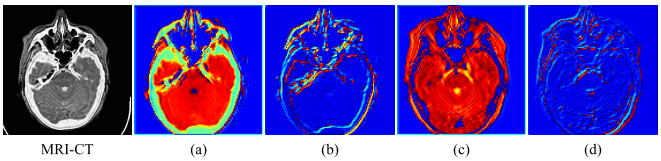

Fig. 11. (a)–(d) Visual illustrations of MRI and CT images from the edge enhancement module to the region mask module. The red areas mean "principal" regions. After passing through the region mask module, "principal" features are learned while these features will no longer be learned in the subsequent process.

location of the edge enhancement module and the region mask module to verify the effect of location on fusion results. The generated results after swapping positions and the corresponding enlarged details are shown in Fig. 12(a).

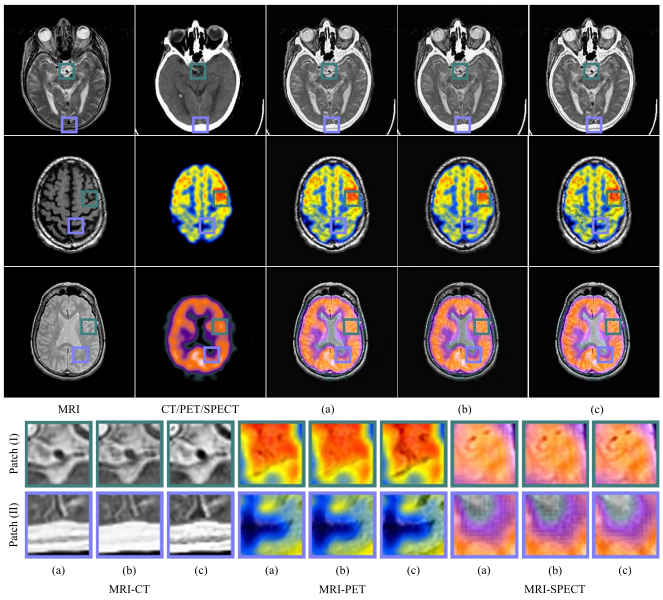

Fig. 12. Ablation qualitative experiment of module position and w/o skip connection operation. Obviously, our generated images contain more texture details and chrominance information. (a) Position exchange, (b) w/o skip connection, and (c) ours.

We can obviously notice that the results present the blur edge details and distorted color. Due to missing edge optimization operations, the region mask module may mistake the edge information as redundant information to prevent the network from recomputing it. Hence, we should use the edge enhancement module to learn edge features first and then deploy the region mask module to mark different regions. Furthermore,

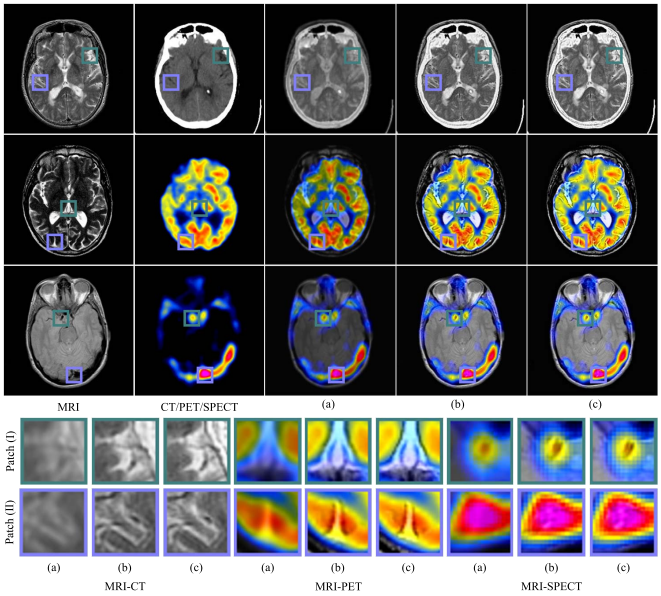

Fig. 13. Ablation qualitative experiment of loss functions. From left to right: (a) results without the edge loss function, (b) results without the semantic loss function, and (c) results with the proposed loss function.

the skip connection operation (referred to as $\mathcal{S}$) from the edge enhancement module to the latter convolution also affects the fusion results on the CT/PET/SPECT branch. With $\mathcal{S}$, we can achieve well-performed fused images, which contain the substantial former learned information. The quantitative results are shown in Fig. 12(b).

3) Analysis on the Loss Function: In Fig. 13, we demonstrate qualitative results of using different combinations of loss functions to train the proposed network. It is easy to find that the results may miss some significant edge details (e.g., the junction of bone and soft tissue in the MRI-CT task) without the edge loss $\mathcal{L}_E$. As a consequence, $\mathcal{L}_E$ has a prominent advantage for edge detail enhancement. When the semantic loss $\mathcal{L}_S$ is removed during training, the semantic information may reduce on fusion images. As shown in Fig. 13(b), undesirable halos are evidently revealed and the performance of color emerges with slight distortion. After integrating all proposed loss functions, we can obtain a fusion result with abundant edge details and semantic information, which helps researchers to understand image contents more conveniently. We give the results of quantitative analysis in Table IV to further verify the effectiveness of each loss function in our proposed method. Evaluation metrics perform well on three fusion tasks, implying that our proposed edge and semantic loss functions are efficient in retaining details and equalizing pixel distribution.

V. Conclusion

In this article, a novel flexible mask-optimized network guided with a semantic model in an end-to-end manner is proposed to conduct the multimodal medical image fusion task, which is named as GeSeNet. Using the edge enhancement module and the corresponding edge loss function, the edge textures of fusion results can be more clear. The region mask module performs improved extraction and skips redundancy operations on different regions after division, while using the global refinement module to modify extracted global features. Furthermore, we employ the semantic module and a newly designed loss function to transfer more semantic information for boosting the quality of fused images. Sufficient experiments show that the proposed GeSeNet model can generate vivid fusion results in visual perception while also guaranteeing the performance of quantitative metrics. Therefore, our proposed method contributes to the development of multimodal medical image fusion. In future works, we will apply the results of multimodal medical image fusion to medical image segmentation and classification, which enables researchers and doctors to judge the disease more accurately.

References

[1] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 49–64, Jan. 2019.

[2] S. Daneshvar and H. Ghassemian, "MRI and PET image fusion by combining IHS and retina-inspired models," *Inf. Fusion*, vol. 11, no. 2, pp. 114–123, Apr. 2010.

[3] W. Li, X. Peng, J. Fu, G. Wang, Y. Huang, and F. Chao, "A multiscale double-branch residual attention network for anatomical–functional medical image fusion," *Comput. Biol. Med.*, vol. 141, Feb. 2022, Art. no. 105005.

[4] B. Huang, F. Yang, M. Yin, X. Mo, and C. Zhong, "A review of multimodal medical image fusion techniques," *Comput. Math. Methods Med.*, vol. 2020, pp. 1–16, Apr. 2020.

[5] W. Tan et al., "Multimodal medical image fusion algorithm in the era of big data," *Neural Comput. Appl.*, 2020, doi: 10.1007/s00521-020-05173-2.

[6] K. He, X. Zhang, D. Xu, J. Gong, and L. Xie, "Fidelity-driven optimization reconstruction and details preserving guided fusion for multi-modality medical image," *IEEE Trans. Multimedia*, early access, Jun. 23, 2022, doi: 10.1109/TMM.2022.3185887.

[7] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 635–648, 2023.

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

[9] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[10] P. Ganasala and A. D. Prasad, "Medical image fusion based on laws of texture energy measures in stationary wavelet transform domain," *Int. J. Imag. Syst. Technol.*, vol. 30, no. 3, pp. 544–557, Sep. 2020.

[11] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[12] 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. 44, no. 1, pp. 502–518, Jan. 2022.

[13] H. Xu and J. Ma, "EMFusion: An unsupervised enhanced medical image fusion network," *Inf. Fusion*, vol. 76, pp. 177–186, Dec. 2021.

[14] Y. Yang, D. S. Park, S. Huang, and N. Rao, "Medical image fusion via an effective wavelet-based approach," *EURASIP J. Adv. Signal Process.*, vol. 2010, no. 1, pp. 1–13, Dec. 2010.

[15] J. Du, W. Li, B. Xiao, and Q. Nawaz, "Union Laplacian pyramid with multiple features for medical image fusion," *Neurocomputing*, vol. 194, pp. 326–339, Jun. 2016.

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

[17] Z. Wang, Z. Cui, and Y. Zhu, "Multi-modal medical image fusion by Laplacian pyramid and adaptive sparse representation," *Comput. Biol. Med.*, vol. 123, Aug. 2020, Art. no. 103823.

[18] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, pp. 3–20, Nov. 2016.
[19] R. Singh, M. Vatsa, and A. Noore, "Multimodal medical image fusion using redundant discrete wavelet transform," in *Proc. 7th Int. Conf. Adv. Pattern Recognit.*, Feb. 2009, pp. 232–235.
[20] O. Prakash and A. Khare, "CT and MR images fusion based on stationary wavelet transform by modulus maxima," in *Computational Vision and Robotics: Proceedings of ICCVR 2014*. India: Springer, 2015, pp. 199–204.
[21] W. Xue-jun and M. Ying, "A medical image fusion algorithm based on lifting wavelet transform," in *Proc. Int. Conf. Artif. Intell. Comput. Intell.*, vol. 3, Oct. 2010, pp. 474–476.
[22] S. Cheng, J. He, and Z. Lv, "Medical image of PET/CT weighted fusion based on wavelet transform," in *Proc. 2nd Int. Conf. Bioinf. Biomed. Eng.*, May 2008, pp. 2523–2525.
[23] V. Bhavana and H. K. Krishnappa, "Multi-modality medical image fusion using discrete wavelet transform," *Proc. Comput. Sci.*, vol. 70, pp. 625–631, Jan. 2015.
[24] A. Sahu, V. Bhateja, A. Krishn, and H. Patel, "Medical image fusion with Laplacian pyramids," in *Proc. Int. Conf. Med. Imag., m-Health Emerg. Commun. Syst. (MedCom)*, Nov. 2014, pp. 448–453.
[25] C. He, Q. Liu, H. Li, and H. Wang, "Multimodal medical image fusion based on IHS and PCA," *Proc. Eng.*, vol. 7, pp. 280–285, Jan. 2010.
[26] A. Krishn, V. Bhateja, H. Patel, and A. Sahu, "Medical image fusion using combination of PCA and wavelet analysis," in *Proc. Int. Conf. Adv. Comput., Commun. Informat. (ICACCI)*, Sep. 2014, pp. 986–991.
[27] Y. Zhu, J. Cheng, Z. Cui, Q. Zhu, L. Ying, and D. Liang, "Physics-driven deep learning methods for fast quantitative magnetic resonance imaging: Performance improvements through integration with deep neural networks," *IEEE Signal Process. Mag.*, vol. 40, no. 2, pp. 116–128, Mar. 2023.
[28] P. Liu, J. Liu, and L. Xiao, "A unified pansharpening method with structure tensor driven spatial consistency and deep plug-and-play priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5413314.
[29] J. Yang, L. Xiao, Y. Zhao, and J. C. Chan, "Variational regularization network with attentive deep prior for hyperspectral–multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5508817.
[30] B. Wen, S. Ravishankar, L. Pfister, and Y. Bresler, "Transform learning for magnetic resonance image reconstruction: From model-based learning to building neural networks," *IEEE Signal Process. Mag.*, vol. 37, no. 1, pp. 41–53, Jan. 2020.
[31] S. Singh and R. S. Anand, "Multimodal medical image fusion using hybrid layer decomposition with CNN-based feature mapping and structural clustering," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 6, pp. 3855–3865, Jun. 2020.
[32] Y. Liu, X. 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.
[33] X. Song, X. J. Wu, and H. Li, "MSDNet for medical image fusion," in *Image and Graphics: 10th International Conference, ICIG 2019, Beijing, China, August 23–25, 2019, Proceedings, Part II 10*. Springer, 2019, pp. 278–288.
[34] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.
[35] R. Liu, J. Liu, Z. Jiang, X. Fan, and Z. Luo, "A bilevel integrated model with data-driven layer ensemble for multi-modality image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 1261–1274, 2021.
[36] Z. Wang and Y. Ma, "Medical image fusion using m-PCNN," *Inf. Fusion*, vol. 9, no. 2, pp. 176–185, Apr. 2008.
[37] X. Xu, D. Shan, G. Wang, and X. Jiang, "Multimodal medical image fusion using PCNN optimized by the QPSO algorithm," *Appl. Soft Comput.*, vol. 46, pp. 588–595, Sep. 2016.
[38] P. Ganasala and V. Kumar, "Feature-motivated simplified adaptive PCNN-based medical image fusion algorithm in NSST domain," *J. Digit. Imag.*, vol. 29, no. 1, pp. 73–85, Feb. 2016.
[39] C. Peng, T. Tian, C. Chen, X. Guo, and J. Ma, "Bilateral attention decoder: A lightweight decoder for real-time semantic segmentation," *Neural Netw.*, vol. 137, pp. 188–199, May 2021.
[40] J. Liu, J. Shang, R. Liu, and X. Fan, "Attention-guided global-local adversarial learning for detail-preserving multi-exposure image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5026–5040, Aug. 2022.
[41] J. Liu et al., "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5802–5811.
[42] 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.
[43] L. Wang et al., "Exploring sparsity in image super-resolution for efficient inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4917–4926.
[44] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[45] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, and L. Van Gool, "Unsupervised semantic segmentation by contrasting object mask proposals," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10052–10062.
[46] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel–Softmax," 2016, *arXiv:1611.01144*.
[47] 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.
[48] W. Tan, W. Thitøn, P. Xiang, and H. Zhou, "Multi-modal brain image fusion based on multi-level edge-preserving filtering," *Biomed. Signal Process. Control*, vol. 64, Feb. 2021, Art. no. 102280.
[49] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
[50] J. Liu, X. Fan, J. Jiang, R. Liu, and Z. Luo, "Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 105–119, Jan. 2022.
[51] H. Li and X. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[52] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 12797–12804.
[53] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, Oct. 2021.
[54] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, p. 313, 2002.
[55] 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, Apr. 2013.
[56] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU, Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, Dec. 2015.
[57] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.

Jiawei Li received the M.S. degree in software engineering from Dalian University, Dalian, China, in 2023.

His current research interests include deep learning and image processing.

Jinyuan Liu (Member, IEEE) received the Ph.D. degree in software engineering from the Dalian University of Technology, Dalian, China, in 2022.

He is currently an Assistant Research Fellow with the School of Mechanical Engineering, Dalian University of Technology. His research interests include computer vision, image processing, and deep learning.

Shihua Zhou (Member, IEEE) was born in Dalian, China, in 1982. She received the Ph.D. degree in mechanical design and theory from the Dalian University of Technology, Dalian, in 2013.

Since 2013, she has been with Dalian University, Dalian, where she is currently an Associate Professor with the Key Laboratory of Advanced Design and Intelligent Computing, Ministry of Education, School of Software Engineering. She has authored more than 50 articles. Her research interests include deoxyribonucleic acid (DNA) computing, DNA self-assembly, image encryption, and image fusion.

Qiang Zhang (Senior Member, IEEE) received the B.S. degree in electronic engineering and the M.S. and Ph.D. degrees in circuits and systems from the School of Electronic Engineering, Xidian University, Xi'an, China, in 1994, 1999, and 2002, respectively.

He is currently the Dean and a Professor with the College of Computer Science and Technology, Dalian University of Technology, Dalian, China. His research interests are artificial intelligence, neural networks, DNA computing, and optimization and intelligent robots.

Nikola K. Kasabov (Life Fellow, IEEE) received the Ph.D. degree from the Technical University of Sofia, Sofia, Bulgaria, in 1975.

He is the Founding Director of Knowledge Engineering and Discovery Research Institute (KEDRI), Auckland, New Zealand, and a Professor of Knowledge Engineering with the School of Engineering, Computing and Mathematical Sciences, Auckland University of Technology, Auckland. He holds the Professorial Chair position with the University of Ulster, Londonderry, U.K., and a Visiting Professorship with the IICT, Bulgarian Academy of Sciences, Sofia, and Dalian University, Dalian, China. He has authored more than 700 articles. His research areas are computational intelligence, neuroinformatics, knowledge discovery, and spiking neural networks.