Stereo Cross-Attention Network for Unregistered Hyperspectral and Multispectral Image Fusion

Yujuan Guo, Xiyou Fu, *Member, IEEE*, Meng Xu, *Member, IEEE*, and Sen Jia, *Senior Member, IEEE*

***Abstract*— The necessary prerequisite for effective data fusion is the strict registration of low-resolution hyperspectral images (LR-HSIs) and high-resolution multispectral images (HR-MSIs). However, registration requires a complex process that takes into account the effects of light, imaging angle, and geometric distortion of the image during acquisition. Therefore, to avoid complex registration, we focused on developing an unregistered HSI and MSI fusion method for pixel shifting, obtaining fused images with high resolution, high signal-to-noise ratio, and feature identifiability. We identified that the unregistered LR-HSI and HR-MSI in the case of pixel shift are very similar to the disparity maps in stereo vision. Inspired by this, we simulate the structure of stereo cameras to propose a stereo cross-attention network (SCANet) to achieve an accurate fusion of unregistered LR-HSI and HR-MSI. Considering the model complexity and computing efficiency, we design a simple and stackable stereo cross-fusion block (SCFBlock) based on a Transformer to simulate the process of light entering the left and right cameras by extracting the abstract features of the images. Moreover, the purpose of cross-convergence fusion self-attention (CCFSA) is to learn cross-complementary attention and collect contextual information in horizontal and vertical directions to fuse unregistered images using multidirectional cross-view information. We have conducted extensive experiments on Pavia University (PaviaU), Chikusei, and PYLake datasets. The results show that the SCANet achieves superior or competitive performance in fusing unregistered LR-HSI and HR-MSI in comparison with the other competitors.**

***Index Terms*— Data fusion, deep learning, hyperspectral image (HSI), registration, stereo cross-attention network (SCANet).**

I. Introduction

DUE to the limitations of the satellite sensor imaging system, the acquired images are mutually constrained in terms of high spatial and hyperspectral resolution and cannot be obtained at the same time [1]. Image fusion can integrate the complementary spatial and spectral advantages of multispectral image (MSI) and hyperspectral image (HSI) to generate a high-resolution HSI (HR-HSI) [2], [3]. HSI and MSI fusion is widely used in remote sensing tasks, such as anomaly detection [4], [5], [6], spatial feature extraction [7], [8], visual image analysis [9], and scene interpretation [10].

Based on the different stages of fusion in the process, MSI and HSI fusion can be divided into pixel level, feature level, and decision level [11]. Pixel-level fusion is generally used for image data with different spectral features, preserving more comprehensive and detailed information in the original image [12], which is very beneficial to image understanding, target detection, recognition, and so on. Therefore, image fusion at the pixel level is the main research content carried out in this article. At present, a large number of fusion algorithms have been proposed, which are mainly classified into traditional and deep learning methods. Traditional fusion methods include component substitution (CS) [13], [14], multiresolution analysis (MRA) [15], [16], Bayesian [17], [18], matrix decomposition [19], [20], [21], and so on. Most of these methods are adapted from pansharpening techniques and achieve good fusion results. However, these methods generally assumed degradation models as a prior [22]. The degradation model can reflect the characteristics of the sensor, but it is not always fully available in the practical applications of remote sensing. Hence, these algorithms need further improvement.

In recent years, deep learning methods have been widely used in the field of computer vision, showing excellent feature extraction capabilities. Meanwhile, some researchers have started to introduce them into the fusion of HSI and MSI [23], [24], [25]. Yang et al. [26] applied deep networks when performing MSI and HSI fusion, which can better extract detailed information from HSIs. Liu et al. [27] proposed a two-stream fusion network (TFNet) to solve the pansharpening problem of MSIs. Zhang et al. [28] proposed a convolutional neural network (CNN)-based spatial–spectral information reconstruction network (SSR-Net) to improve the spatial resolution of fused HSIs. Realizing that CNNs have difficulty in capturing long-term dependencies in images, the Transformer-based model aims at modeling remote dependencies through a self-attentive mechanism [29]. Hu et al. [30] designed a Transformer-based architecture (called Fusformer), which can globally explore the intrinsic relationship within features. The advantage of deep learning methods is that all parameters in the network can be updated under the supervision of training samples, thus reducing the need for prior knowledge, and higher fitting accuracy can be expected. Therefore, deep learning is gradually becoming the mainstream method for image fusion.

However, both traditional and deep learning fusion methods are designed based on the premise of strict registration of HSI and MSI [31]. In general, two images of the same scene

Manuscript received 8 June 2023; revised 26 July 2023; accepted 6 September 2023. Date of publication 13 September 2023; date of current version 25 September 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62271327 and Grant 41971300; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515011290 and Grant 2022A1515110076; and in part by the Shenzhen Science and Technology Program under Grant RCJC20221008092731042, Grant JCYJ20220818100206015, and Grant KQTD20200909113951005. *(Corresponding author: Sen Jia.)*

The authors are with the College of Computer Science and Software Engineering, the Guangdong–Hong Kong–Macau Joint Laboratory for Smart Cities, and the Key Laboratory for Geo-Environmental Monitoring of Coastal Zone, Ministry of Natural Resources, Shenzhen University, Shenzhen 518060, China (e-mail: guoyujuan@szu.edu.cn; fuxiyou@qq.com; m.xu@szu.edu.cn; senjia@szu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3314755

acquired by different sensors with different viewpoints can cause geometric distortions, such as squeezing, stretching, distortion, and translation due to illumination, season, angle, and other factors [32]. In this case, image registration is required to convert the images to the same coordinate system to eliminate the geometric errors between them. Therefore, registration is an essential step in the current image fusion process. However, regardless of the registration methods adopted, mismatching points will always be generated due to nonlinear gray distortion and strong noise interference [33], which will undoubtedly increase the computational burden and manual involvement. Currently, robust registration algorithms for large-scale multisource images are not available. Image fusion can attenuate the modal differences of multisource data and reduce the impact of redundant information on the registration process. Therefore, the development of complementary robust algorithms for registration-image fusion is expected in fusion scenarios with relatively large modal differences.

With the above considerations, it is meaningful for us to design deep learning fusion methods for unregistered HSI and MSI. Inspired by the stereo vision imaging process [34], [35], we noticed that the unregistered HSI and MSI are similar to the scenes recorded by stereo cameras (temporarily disregarding the presence of squeezed, stretched, and distorted HSI and MSI). Therefore, we propose a stereo cross-attention network (SCANet) based on a Transformer, which consists of a two-branch parallel weight-sharing network to simulate the stereo camera structure. A stereo cross-fusion block (SCFBlock) is designed in SCANet to control the information flow-through channel attention and gate units to gradually pass the fusion features from shallow to deep layers. Furthermore, the construction of cross-convergence fusion self-attention (CCFSA) makes SCANet powerful in cross-view information exploitation. It should be noted that the differences between unregistered images vary significantly with increasing errors, which poses a significant challenge to capturing a reliable correspondence between pixels. Finally, we design a simple fusion module to exploit the global spatial information. We conducted extensive quantitative and qualitative experiments on several datasets to demonstrate the effectiveness of our proposed SCANet.

The contributions of this article are as follows.

1) The proposed SCANet can effectively avoid the complicated registration process, and it aims to improve the resolution, information content, sharpness, and signal-to-noise ratio of the fused images and enhance the identifiability of features in the fused images.
2) We designed a simple and efficient SCFBlock that takes as a reference and simplifies the module constructed by Restormer [36]. SCFBlock improves computational efficiency and controls information flow by simplifying the channel attention mechanism and gate unit to capture global information while focusing on different information details.
3) Considering that features at different locations have different importance to the fusion task, this article proposes CCFSA. CCFSA can focus on the correlation between the target pixels and the remaining pixels, so that pixels at different locations have the same chance of expression and, thus, capture the rich background information. CCFSA is performed with complementary features generated by SCFBlock for cross-view interaction. The unregistered images are fused using cross-view information in multiple directions by collecting contextual information in horizontal and vertical directions.

SCANet is a rewarding attempt at a pixel-level fusion method for unregistered HSI and MSI. It provides a feasible and beneficial attempt to reduce the data preprocessing process and improve the efficiency of remote sensing applications.

The remainder of this article is organized as follows. Section II describes the related work on HSI and MSI fusion. Section III describes in detail the principles of stereo vision. Section IV describes the architecture of SCANet and elaborates SCFBlock, CCFSA, and global residual feature fusion modules (GRFMs). Section V provides an in-depth discussion of SCANet through ablation experiments and analyzes the experimental results of the comparison method on the Pavia University (PaviaU), Chikusei, and PYLake datasets. Section VI draws comprehensive conclusions and provides an outlook on possible future research directions.

II. Relate Work

The fusion of low-resolution HSIs (LR-HSIs) and high-resolution MSIs (HR-MSIs) can be approximately divided into two categories: traditional methods and deep learning methods. Therefore, this article presents related work from the below two aspects.

A. Traditional Methods

Currently, the traditional fusion methods of LR-HSI and HR-MSI mainly include CS, MRA, model optimization, and matrix decomposition methods.

The CS [37], [38] and MRA [39], [40] methods were first designed for remote sensing image pansharpening, and their applications can be extended to HSI and MSI fusion. The model optimization approach treats fusion as an inverse problem. It models the relationship between the HR-HSI to be fused and the LR-HSI and HR-MSI based on the degradation mechanism of the spectral image and solves it by using an optimization algorithm to obtain the fused image [41], [42]. The basic idea based on matrix decomposition is to decompose the original matrix into the product of two matrices, thus reducing the original high-dimensional matrix [43]. The matrix decomposition is used to reduce the dimensionality of the original image, and the decomposition results have a more explicit physical meaning. Typical algorithms contain NMF [44] and coupled NMF (CNMF) [45]. Although these methods have achieved good results in HSI and MSI fusion, shortcomings still exist. The CS methods can lead to the distortion of the fused image spectral due to the incomplete wavelength coverage of the two images. The MRA methods only extract the high-frequency part of the high-resolution image, and the fused image faces the problem of insufficient spatial resolution improvement. Compared with the first two methods, the fusion methods based on model optimization have higher fusion

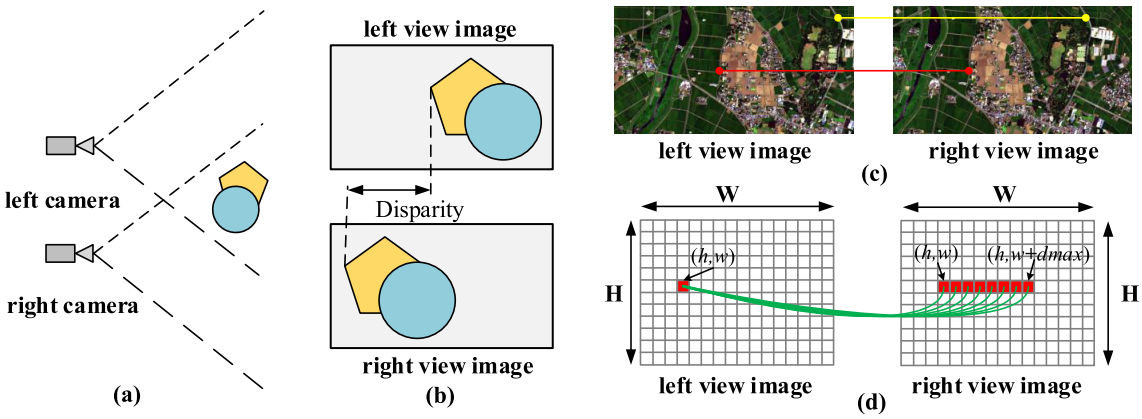

Fig. 1. Basic principles of the stereo image. (a) Simulation diagram of the stereo camera. (b) Images recorded by stereo cameras. (c) Unregistered HSI and MSI. (d) Epipolar constraint relationship for stereo images.

accuracy, but the model solution is complicated. In addition, the model optimization-based fusion method relies seriously on manually designed features.

The way multiple remote sensing data sources are imaged leads to their intrinsic relationships being more complex. It is difficult for traditional methods to utilize these features in an integrated way. Traditional feature extraction methods destroy the original spatial–spectral structure in the image, thus ignoring a large amount of implicit valid information. Therefore, exploring more suitable feature extraction methods is an important research direction for data fusion.

B. Deep Learning Methods

Compared with traditional methods, deep learning is broadly used in remote sensing image processing and analysis for its end-to-end integrity training approach and effective abstract feature mining.

Currently, deep learning fusion methods are predicated on strict registration [30]. Although some models [46] are designed for the pansharpening problem, they can also be directly applied to the fusion of HSI and MSI. He et al. [47] proposed a hyperspectral pansharpening neural network (HyperPNN) for the fusion of HSI and panchromatic images, which improves the spectral prediction capability of the network by adding a spectral prediction layer. Palsson et al. [48] designed a 3DCNN network for HSI and MSI fusion, using 3-D convolution to extract features of the input image. However, all of the above methods considered both images as a whole, and the network input was a merging of the two images along the channel dimension, ignoring their respective significant characteristics. To solve this problem, more and more scholars have abandoned the use of single-branch networks. For instance, a remote sensing image fusion neural network (RSIFNN) containing two branches was designed by Shao and Cai [49] to extract the features of MSI and panchromatic images, respectively, and fused the two features for image reconstruction. Jiamin and Huihui [50], inspired by U-net, divided the network into an encoding–decoding structure, where the feature extraction part used two subnetworks to extract the features of MSI and panchromatic image, respectively. Zhou et al. [51] designed a two-stream network with the self-attention to extract the modality-specific features from the panchromatic images and MSI modalities and apply a cross-attention module to merge the spectral and spatial features. This method also pays attention to the influence of cross-complementary information on image fusion but does not take into account the importance of multidirectional and long-range feature representation. The above methods of CNN-based deep learning exhibit excellent feature extraction capabilities. CNNs can extract multiple nonlinear features with high invariance from hyperspectral data by convolutional operations and mine complex relationships with disparate features in multisource data.

The state-of-the-art approaches have employed CNNs to encode meaningful features for image fusion [25], [27]. However, they do not consider the long-term dependencies in the images. In recent years, more researchers have focused on Transformer models, aiming to overcome this by modeling remote dependencies with the help of self-attention mechanisms [30]. Vs et al. [52] developed a Transformer-based multiscale fusion strategy that simultaneously processes local and remote information for the fusion of infrared and visible images. Qu et al. [53] proposed a Transformer-based multiexposure image fusion framework (TransMEF) using self-supervised multitask learning. The framework is based on an encoder–decoder network and can be trained on large natural image datasets. Wang et al. [54] proposed a new multilevel cross-transformation algorithm (MCT-Net) to obtain the global contextual information of two images to achieve sufficient fusion of spectral and spatial information. Transformer makes few assumptions about structural deviations in the input data and is, therefore, a very flexible and versatile architecture.

The current fusion model is designed based on the premise of strict image registration. Therefore, the development of robust registration-fusion integration algorithms is expected in fusion scenarios with significant geometric errors. Guo et al. [55] first constructed a CNN, called RegNet, to register pixel-level offsets between HSI and MSI. Zhou et al. [56] proposed a registration algorithm that incorporates a point spread function (PSF) into a minimization least square (LSQ) objective function applied to the fusion

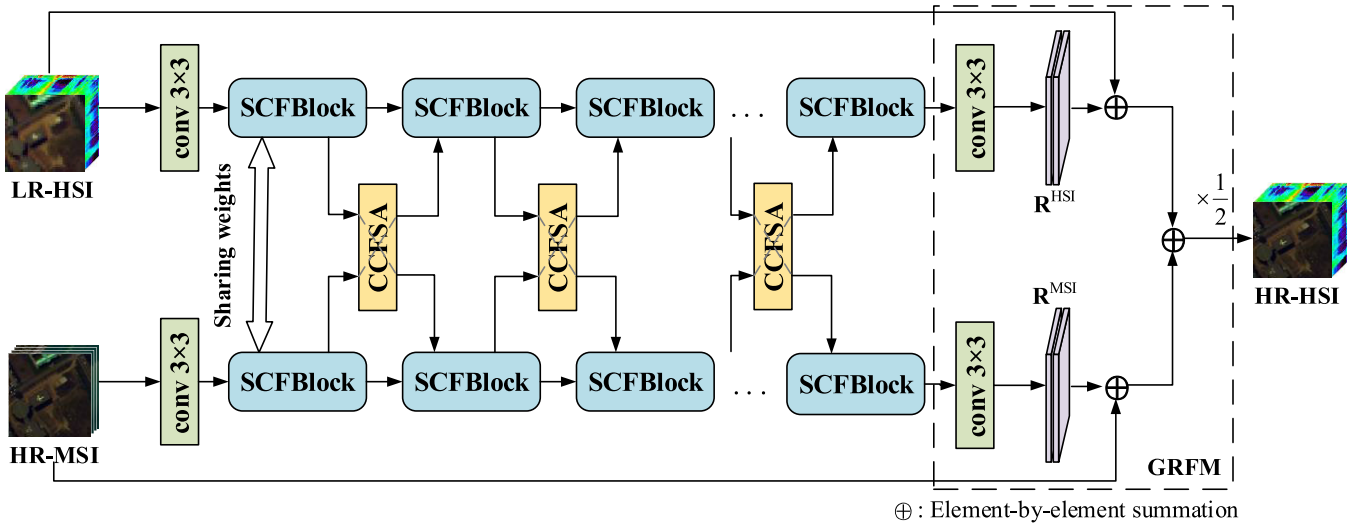

Fig. 2. Overall structure of SCANet. SCFBlock is a stereo cross-fusion block. CCFSA is a cross-convergence fusion self-attention module. GRFM is a global residual feature fusion module.

of HSI and MSI. Nie et al. [57] learned the reconstruction of the unregistered HSI with affine transform parameters between the input two images by introducing a spatial Transformer network (STN). Fu et al. [58] systematically evaluated the effects of different methods for geometric misalignment on RGB and HSI fusion. Qu et al. [31] proposed an unregistered and unsupervised mutual Dirichlet-Net (u2-MDN), which does not require multimodal registration to solve the HSI super-resolution problem. Zheng et al. [59] proposed a novel unsupervised spectral unmixing and image deformation correction network, NonRegSRNet, which integrates dense registration and super-resolution tasks into a unified model. The above approach tries to avoid the image registration process by designing a "registration + fusion" connection network, but automatic and high-precision image registration remains a challenging problem. Moreover, current fusion methods mainly focus on the simple superposition of feature modules, ignoring the intrinsic connection between multisource remote sensing data. Therefore, the proposed SCANet in this article avoids the design of the registration algorithm as a meaningful attempt.

III. Basic Principles of Stereo Vision

The stereo camera uses the left and right views to record the current scene. The complementary information between the left and right views is cross-referenced during the fusion process, which provides an additional constraint for image fusion. Fig. 1(a) shows a schematic of the stereo camera imaging model. Two cameras with parallel optical axes capture the target scene from the left and right viewpoints, and the recorded images are shown in Fig. 1(b). The left and right views of the stereoscopic camera are offset, so the same object in the scene is positioned somewhat differently in the left and right views. The imaging process is very similar to that of unregistered HSI and MSI (Fig. 1(c), only the case of translation is considered).

In the stereo imaging process, a stereo camera with the same focal length, the same pointing (i.e., parallel to the optical axis), and two camera lines (baselines) perpendicular to the optical axis is usually used. Under these conditions, only horizontal parallax exists for the same object in the left- and right-view images. The relationship between parallax and the scene depth (the distance between the object and the camera) can be expressed as follows:

$$\gamma = \frac{Bf}{d} \tag{1}$$

where γ is the scene depth; B is the baseline length between the left and right cameras; f is the focal length of the camera, and d is the parallax of the object in the left-view and right-view images.

Using the complementary information between the left and right views to improve the reconstruction quality of stereo images, it is necessary to establish associations in the corresponding regions of the left- and right-view images. Without considering the occlusion, the corresponding regions of the left- and right-view images should be on the same height horizontal line, which is called the polar line. The fusion algorithm for unregistered images should establish the association between the left- and right-view images in conjunction with the polar line constraint. The limit constraint relationship of stereo images is shown in Fig. 1(d); that is, for any point (h, w) in the left-view image, its corresponding point in the right-view image should lie between (h, w) and $(h, w + d_{\max})$, where w and h represent the width and height of the left- and right-view images, respectively, and $d_{\max}$ is the maximum value of the parallax between the left- and right-view images.

IV. Method

In this section, we give the details of the proposed SCANet for the fusion of images that are not strictly registered. We first present the general framework of our SCANet. We then detail our SCFBlock and the CCFSA. Finally, we propose a simple fusion module to exploit the global spatial information.

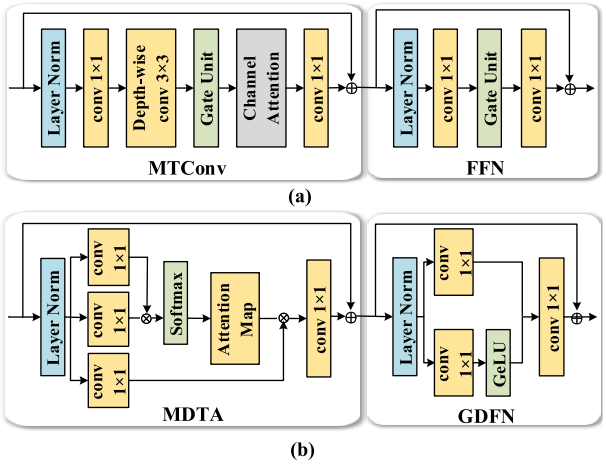

Fig. 3. (a) SCFBlock consists of two submodules: MTConv module and FFN. (b) Restormer [36] consists of two submodules: MDTA and GDFN.

A. Overall Framework

The overall pipeline of our proposed SCANet is shown in Fig. 2. Inspired by stereo vision imaging [60], SCANet consists of a two-branch parallel weight-sharing network to simulate the stereo camera structure. SCANet consists of three main components: the SCFBlock module, the CCFSA module, and a GRFM.

SCANet takes LR-HSI and HR-MSI as input, respectively. HSI and MSI are first mapped to feature spaces of the same dimension by a 3 × 3 convolutional layer, respectively. Then, the image features are extracted, in turn, by n SCFBlocks. In addition, to increase the information interaction between LR-HSI and HR-MSI, a CCFSA is inserted after each SCFBlock for weight sharing and feature fusion. The CCFSA performs bidirectional cross-view interaction by combining complementary features generated by SCFBlock. The full utilization of LR-HSI and HR-MSI contextual information is achieved by fusing the interaction information with the input image features. The details of SCFBlock and CCFSA are shown in Figs. 3(a) and 4. Finally, to reduce the burden of feature learning, a simple GRFM is used to implement the fusion of LR-HSI and HR-MSI. Compared with other two-branch fusion networks, SCANet achieves geometric registration of HSI and MSI by simulating stereo vision structures and then learns and matches attributes between different modal information through parameter sharing, which allows SCANet to avoid complex feature fusion operations.

B. Stereo Cross-Fusion Block

Restormer [36] is an efficient Transformer that enables it to capture long-range pixel interactions while still being suitable for high-resolution images by performing several key designs in the building blocks [multi-Dconv head transposed attention (MDTA) and gated-Dconv feed-forward network (GDFN), Fig. 3(b)]. Therefore, based on the consideration of model complexity and computational efficiency, the SCFBlock is designed to be simplified with Restormer as a reference. By superimposing multiple SCFBlocks, the deeper features of the image can be extracted gradually to simulate the process of light entering the left and right cameras.

The details of SCFBlock are shown in Fig. 3. SCFBlock consists of two submodules: the multitranslation convolution (MTConv) module and the feed-forward network (FFN) module. Both modules use residual connections. The whole process is formulated as follows:

$$\begin{aligned}\mathbf{X} &= \mathrm{MTConv}(\mathrm{LN}(\mathbf{X})) + \mathbf{X}\\ \mathbf{X} &= \mathrm{FFN}(\mathrm{LN}(\mathbf{X})) + \mathbf{X}.\end{aligned} \tag{2}$$

MTConv goes through a normalization layer, a 1 × 1 convolution layer, a 3 × 3 depthwise convolution layer, a gate unit, simplified channel attention, and a 1 × 1 convolution layer. Formally, given an input $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$ (H, W, and C are the height, width, and the number of channels, respectively), the MTConv is represented as follows:

$$\begin{aligned}\mathrm{MTConv}(\mathbf{X}) &= (N_1 \circ D \circ G \circ S \circ N_2)(\mathrm{LN}(\mathbf{X})) + \mathbf{X}\\ N_1(\mathbf{X}) &= \mathrm{Conv}(\mathbf{X})\\ D(\mathbf{X}) &= \mathrm{DepthConv}(\mathbf{X})\\ G(\mathbf{X}) &= \mathrm{Gating}(\mathrm{GeLU}(\mathbf{X}))\\ S(\mathbf{X}) &= \mathrm{SCA}(\mathbf{X})\\ N_2(\mathbf{X}) &= \mathrm{Conv}(\mathbf{X}).\end{aligned} \tag{3}$$

Among them, LN and GeLU denote layer normalization and Gaussian error linear unit, respectively. The MTConv b1ock consists of five main functions: N_1, D, G, S, and N_2, corresponding to a 1 × 1 convolution, a 3 × 3 depthwise convolution, a gate unit, simple channel attention, and a 1 × 1 convolution, respectively. Compared with Restormer, we use 3 × 3 depthwise convolution and gate unit instead of self-attention and nonlinear activation functions (e.g., rectified linear unit (ReLU) and GeLU), respectively, making the module more concise and efficient [Fig. 3(a)]. A 3 × 3 depthwise convolution can learn local structural information and is simpler than self-attention [61]. The gate unit first divides the input $\mathbf{X} \in \mathbf{R}^{H\times W\times C}$ into two subfeatures $\mathbf{X}_1, \mathbf{X}_2 \in \mathbf{R}^{H\times W\times C/2}$ along the channel dimension and multiplies them together. The gate unit is expressed as follows:

$$\mathrm{Gating}(\mathbf{X}) = \mathbf{X_1} \otimes \mathbf{X_2} \tag{4}$$

where ⊗ stands for element-by-element multiplication. In addition, related studies have shown that channel attention can meet the computational efficiency requirements and capture global information [62]. Therefore, we further add the channel attention after the gate unit and remove the ReLU, sigmoid, and 1 × 1 convolution layers of the regular channel attention. The simplified expression of channel attention is as follows:

$$\mathrm{SCA}(\mathbf{X}) = \mathbf{X} * W * \mathrm{pool}(\mathbf{X}) \tag{5}$$

where ∗ is a channelwise product operation; W denotes the learnable matrix, and pool is the global average pooling operation.

FFN goes through a normalization layer, a 1 × 1 convolution layer, a gate unit, and a 1 × 1 convolution layer. The

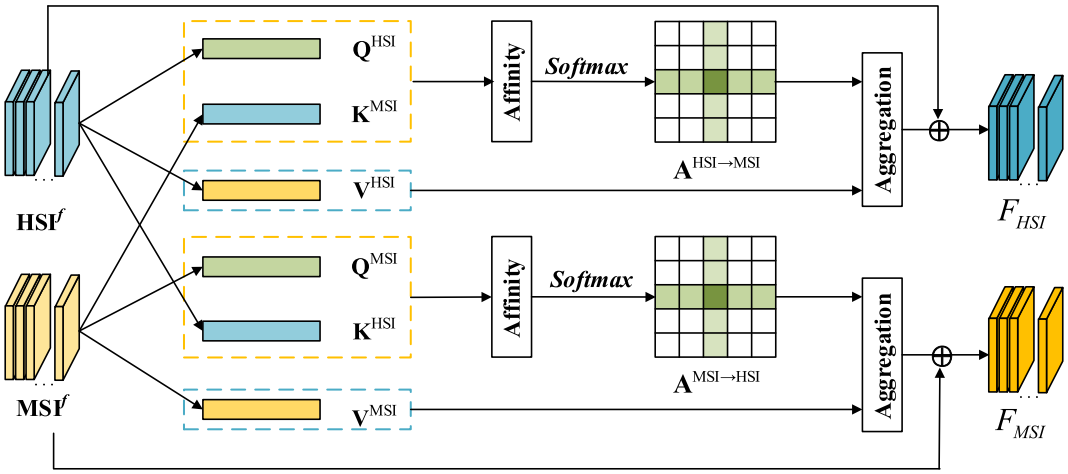

Fig. 4. Structure of CCFSA.

processing of FFN can be expressed as follows:

$$\begin{aligned} \mathrm{FFN}(\mathbf{X}) &= (N_1 \circ G \circ N_2)(\mathbf{LN}(\mathbf{X})) + \mathbf{X} \\ N_1(\mathbf{X}) &= \mathrm{Conv}(\mathbf{X}) \\ G(\mathbf{X}) &= \mathrm{Gating}(\mathrm{GeLU}(\mathbf{X})) \\ N_2(\mathbf{X}) &= \mathrm{Conv}(\mathbf{X}). \end{aligned} \tag{6}$$

In summary, FFN that introduces a gate unit can control the flow of information, thus allowing each layer to focus on different details of the information.

C. Cross-Convergence Fusion Self-Attention

For unregistered images, features at different locations are of different importance for the fusion task. CCFSA is to capture rich contextual information by taking into account the correlation between the target pixel and the rest of the pixels, so that pixels at different distances have the same chance of expression. CCFSA is to use the complementary features generated by SCFBlock as input for cross-view interaction. The purpose of CCFSA is to learn cross-complementary attention and collect contextual information in horizontal and vertical directions to fuse unregistered images using multidirectional cross-view information. CCFSA can interact symmetrically and compactly for the formation of fused images in both directions.

The structure of CCFSA is shown in Fig. 4. The feature maps generated after a two-branch parallel SCFBlock are $\mathbf{HSI}^f, \mathbf{MSI}^f \in \mathbf{R}^{H\times W\times C}$, respectively ($H$, W, and C are the height, width, and the number of channels, respectively). Three 1×1 convolutions are used to generate the query matrix $\mathbf{Q}^{\mathrm{HSI}}, \mathbf{Q}^{\mathrm{MSI}} \in \mathbf{R}^{H\times W\times C}$, the key matrix $\mathbf{K}^{\mathrm{HSI}}, \mathbf{K}^{\mathrm{MSI}} \in \mathbf{R}^{H\times W\times C}$, and the value matrix $\mathbf{V}^{\mathrm{HSI}}, \mathbf{V}^{\mathrm{MSI}} \in \mathbf{R}^{H\times W\times C}$ for $\mathbf{HSI}^f$ and $\mathbf{MSI}^f$, respectively.

We calculate the cross correlation between each pair of query matrix $\mathbf{Q}$ and key matrix $\mathbf{K}$ and apply the softmax function to obtain the corresponding attention weights. Take the example of computing the cross correlation between $\mathbf{Q}^{\mathrm{HSI}}$ and $\mathbf{K}^{\mathrm{MSI}}$. Let $\mathbf{Q}_u^{\mathrm{HSI}} \in \mathbf{R}^{1\times C}$ be the channel dimensional feature vector of $\mathbf{Q}^{\mathrm{HSI}}$ at position u, whose size is $1 \times C$. Accordingly, obtain the set of key vectors $\mathbf{K}_u^{\mathrm{MSI}} \in \mathbf{R}^{(H+W-1)\times C}$ in $\mathbf{K}^{\mathrm{MSI}}$ centered on u, in the row and column directions. Since there are a total of $(H+W-1)$ positions (crosses), the vector size is $(H+W-1) \times C$. $\mathbf{K}_{i,u}^{\mathrm{MSI}} \in \mathbf{R}$ is the ith element of $\mathbf{K}_u^{\mathrm{MSI}}$, $i \in \{1 \ldots, H+W-1\}$. The cross-correlation degree of $\mathbf{Q}_u^{\mathrm{HSI}}$ and $\mathbf{K}_{i,u}^{\mathrm{MSI}}$ is calculated by **affinity** operation as follows:

$$\begin{aligned} d_{i,u}^{\mathrm{HSI}\to\mathrm{MSI}} &= \mathbf{Q}_u^{\mathrm{HSI}} \left(\mathbf{K}_{i,u}^{\mathrm{MSI}}\right)^T \\ d_{i,u}^{\mathrm{MSI}\to\mathrm{HSI}} &= \mathbf{Q}_u^{\mathrm{MSI}} \left(\mathbf{K}_{i,u}^{\mathrm{HSI}}\right)^T \end{aligned} \tag{7}$$

$$\begin{aligned} \mathbf{A}^{\mathrm{HSI}\to\mathrm{MSI}} &= \mathrm{Softmax}\left(\frac{d_{i,u}^{\mathrm{HSI}\to\mathrm{MSI}}}{\sqrt{C}}\right) \\ \mathbf{A}^{\mathrm{MSI}\to\mathrm{HSI}} &= \mathrm{Softmax}\left(\frac{d_{i,u}^{\mathrm{MSI}\to\mathrm{HSI}}}{\sqrt{C}}\right) \end{aligned} \tag{8}$$

where T is the matrix transpose and $d_{i,u}^{\mathrm{HSI}\to\mathrm{MSI}} \in \mathbf{R}^{(H+W-1)\times W\times H}$. After that, the softmax is used in the $(H+W-1)$ dimension to obtain the attention response graph $\mathbf{A}^{\mathrm{HSI}\to\mathrm{MSI}} \in \mathbf{R}^{(H+W-1)\times W\times H}$. The same procedure is used for the calculation between $\mathbf{Q}^{\mathrm{MSI}}$ and $\mathbf{K}^{\mathrm{HSI}}$.

To aggregate the information, the obtained attention graph $\mathbf{A}$ and the value matrix $\mathbf{V}$ are reweighted by the **aggregation** operation for features. Similar to the above process, the value matrix $\mathbf{V}^{\mathrm{HSI}}$ is centered at the point u, and the set of value vectors $\mathbf{V}_u^{\mathrm{HSI}} \in \mathbf{R}^{(H+W-1)\times C}$ is obtained in the row and column directions, and **aggregation** is defined in Fig. 4 as follows:

$$\begin{aligned} F_u^{\mathrm{HSI}\to\mathrm{MSI}} &= \sum_{i\in\left|\mathbf{V}_u^{\mathrm{HSI}}\right|} \mathbf{A}_{i,u}^{\mathrm{HSI}\to\mathrm{MSI}} \mathbf{V}_{i,u}^{\mathrm{HSI}} \\ F_u^{\mathrm{MSI}\to\mathrm{HSI}} &= \sum_{i\in\left|\mathbf{V}_u^{\mathrm{MSI}}\right|} \mathbf{A}_{i,u}^{\mathrm{MSI}\to\mathrm{HSI}} \mathbf{V}_{i,u}^{\mathrm{MSI}} \end{aligned} \tag{9}$$

where the size of $F_u^{\mathrm{HSI}\to\mathrm{MSI}}$ is $H \times W \times C$; $\mathbf{A}_{i,u}^{\mathrm{HSI}\to\mathrm{MSI}}$ is a scalar, which is the ith feature vector corresponding to the attention response graph $\mathbf{A}^{\mathrm{HSI}\to\mathrm{MSI}}$ at position u; and $\mathbf{V}_{i,u}^{\mathrm{HSI}}$ is the ith feature vector in the set $\mathbf{V}_u^{\mathrm{HSI}}$. $F_u^{\mathrm{MSI}\to\mathrm{HSI}}$ is calculated in the same way. In this way, we capture the remote contextual information in the horizontal and vertical directions at position u in the interaction feature.

Finally, the interacting cross-attention information $F_u^{\mathrm{HSI}\to\mathrm{MSI}}$ and $F_u^{\mathrm{MSI}\to\mathrm{HSI}}$ and the information $\mathbf{HSI}^f$ and $\mathbf{MSI}^f$ within the double branch are fused by elemental

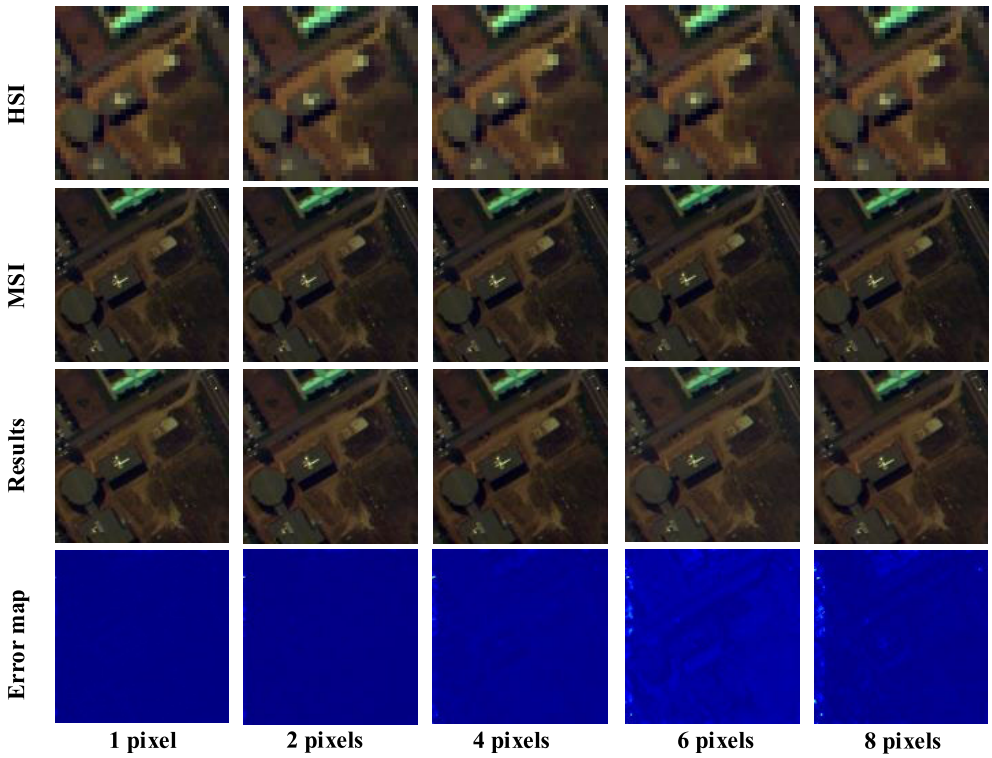

Fig. 5. Fusion results of SCANet in the horizontal registration error direction on the PaviaU dataset.

addition

$$F_{\mathrm{HSI}} = \gamma_{\mathrm{HSI}} F^{\mathrm{HSI}\to\mathrm{MSI}} + \mathbf{HSI}^f$$
$$F_{\mathrm{MSI}} = \gamma_{\mathrm{MSI}} F^{\mathrm{MSI}\to\mathrm{HSI}} + \mathbf{MSI}^f \tag{10}$$

where γ_{HSI} and γ_{MSI} are trainable channelwise scales and initialized with zeros for stabilizing training. The computational complexity of the CCFSA is only $O[(H + W - 1) \times HW]$, which can be lightly embedded in the feature extraction module.

D. Global Residual Feature Fusion Module

A simple GRFM (Fig. 2) is located at the end of the SCANet. A 3×3 convolution is applied to the refined features to produce the residual images $\mathbf{R}^{\mathrm{HSI}} \in \mathbf{R}^{H\times W\times C}$ and $\mathbf{R}^{\mathrm{MSI}} \in \mathbf{R}^{H\times W\times C}$, respectively, and the image inputs are summed to obtain the reconstructed images, i.e., $\mathbf{I}^{\mathrm{HSI}} = \mathbf{X}^{\mathrm{HSI}} + \mathbf{R}^{\mathrm{HSI}}$ and $\mathbf{I}^{\mathrm{MSI}} = \mathbf{X}^{\mathrm{MSI}} + \mathbf{R}^{\mathrm{MSI}}$. Finally, the images of the upper and lower branches are summed and fused to obtain the HR-HSI, i.e., $\mathbf{I} = (\mathbf{I}^{\mathrm{HSI}} + \mathbf{I}^{\mathrm{MSI}})/2$.

V. Experiments

To evaluate the effectiveness of the SCANet, we carry out comprehensive experiments on the PaviaU, Chikusei, and PYLake datasets. Experimental results show that the SCANet achieves state-of-the-art performance on the PaviaU, Chikusei, and PYLake datasets, where HSI and MSI are not strictly registered. In Sections V-A–V-D, we first introduce the datasets and implementation details; then, we perform a series of ablation experiments on the PaviaU dataset. Finally, we report the results of comparison methods on the Chikusei and PYLake datasets.

A. Implementation Details

1) Evaluation Metrics: Root mean square error (RMSE), peak signal-to-noise ratio (PSNR), spectral angle mapper (SAM), structural similarity (SSIM), and erreur relative globale adimensionnelle de synthèse (ERGAS) are used to evaluate the objective evaluation metrics of image fusion methods.

2) Experimental Data: We use the PaviaU, Chikusei, and PYLake datasets to verify the validity of SCANet.

For PaviaU and Chikusei datasets, the LR-HSI is simulated by using a 5 × 5 Gaussian filter with a standard deviation of 2 and then by downsampling with the ratio of r from the reference HR-HSI. The HR-MSI of five bands is generated by a Landsat 8 spectral reflectance response (SRF) filter. The blue–green–red bands of the Landsat 8 SRF were used for PaviaU and Chikusei. In addition, a subarea of 800 × 800 pixels in the Chikusei dataset was cropped for the experiment to reduce the computer's operational burden.

The PYLake dataset is located at Poyang Lake, China. The PYLake dataset contains Sentinel-2A MSIs with a spatial resolution of 10 m and a size of 1200 × 1200 pixels in four bands. The GF-5 hyperspectral data have a spatial resolution of 30 m, a size of 400 × 400 pixels, and a total of 284 bands after removing water vapor absorption and heavy noise bands with a band range of 0.4–2.5 μm and serve as a reference HR-HSI with an acquisition time close to that of Sentinel-2A. The PYLake dataset was preprocessed with the environment for visualizing images (ENVI). First, the multispectral and hyperspectral data were orthorectified using the rational polynomial coefficient (RPC) orthorectification module. Then, the MSI is the reference image, and the HSI is the image to be registered to select the control points with the same name for registration, and the registration error is less than 0.5 pixels.

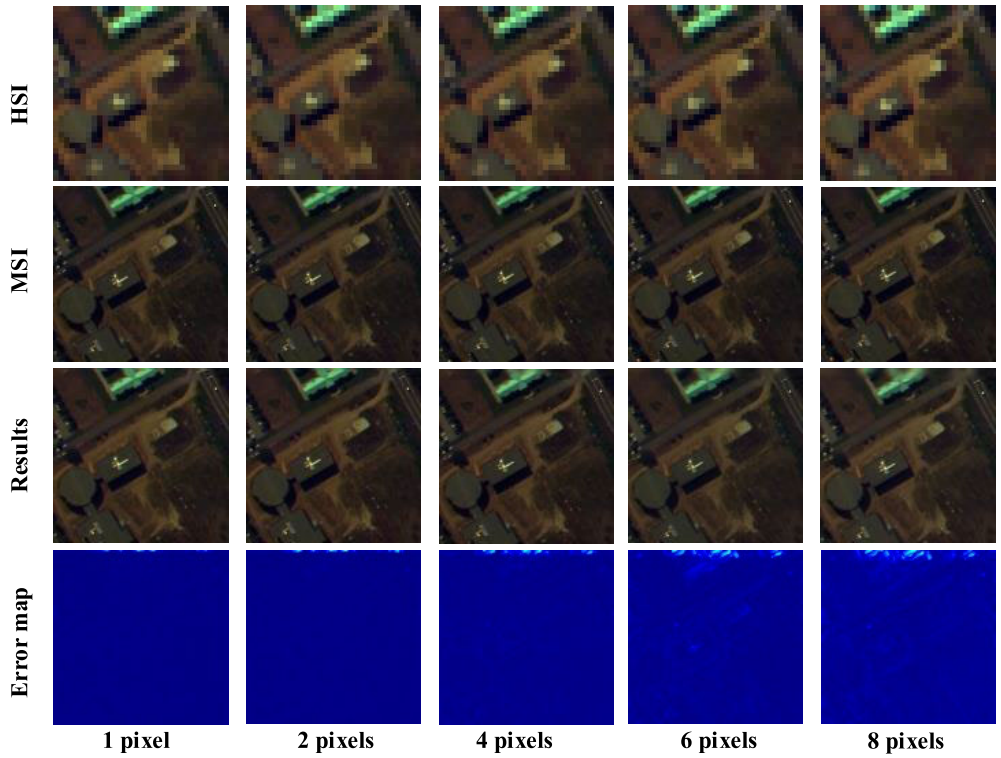

Fig. 6. Fusion results of SCANet in the vertical registration error direction on the PaviaU dataset.

TABLE I

4× FUSION RESULTS (PSNR AND SSIM) ACHIEVED ON THE PAVIAU DATASET BY SCANET WITH DIFFERENT NUMBERS OF SCFBLOCKS. *H* AND *D* REPRESENT THE REGISTERED ERRORS IN THE HORIZONTAL AND DIAGONAL ERROR DIRECTIONS, RESPECTIVELY. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

The number of Blocks	***n*=2**		***n*=4**		***n*=8**		***n*=16**		***n*=32**	
Error directions	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**	**H**	**D**
PSNR↑	39.79	37.67	40.84	37.78	41.08	**38.36**	41.22	38.1	**41.25**	38.03
SSIM↑	0.914	0.915	0.921	0.920	0.923	0.922	0.922	**0.924**	**0.924**	0.923
SAM↓	2.249	2.181	2.041	2.098	2.016	2.034	1.991	**1.980**	**1.935**	2.004
Model Size(M)	**1.66**		1.87		2.28		3.11		4.77	

The LR-HSI is simulated by using a 5 × 5 Gaussian filter with a standard deviation of 2 and then by downsampling with the ratio of r from the reference GF-5. The HR-MSI is simulated by downsampling from Sentinel-2A to 30 m using bilinear interpolation.

3) Training: Hyperspectral and multispectral unregistered states are simulated by translating i pixels in horizontal, vertical, and diagonal directions. All models are optimized by the Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with weight decay 0 by default. The learning rate is 1×10^{-3}. MSE is the loss function. PaviaU and PYLake data are cropped with a 128 × 128 pixels subregion in the center as the test image, and the rest of the area is used for training; 256 × 256 pixels subregion in the center of Chikusei is cropped as the test image, and the rest of the area is used for training.

Besides, all the experiments are implemented by PyTorch 1.10.0 on Python 3.7. The model was trained on a PC with 3.1-GHz eight-core CPUs and 32-GB memory. The NVIDIA NVS 310 GPU was used for acceleration.

B. Ablation Study

To verify the rationality of the SCANet, we conduct extensive ablation experiments on the PaviaU dataset with different settings for SCANet.

1) Number of SCFBlocks: We analyze the effect of different sizes of SCANet on the fusion effect of unregistered images (two pixels error) with different error directions. We construct five different sizes of SCANet by adjusting the number of SCFBlock, which are named SCANet-T ($n = 2$), SCANet-S ($n = 4$), SCANet-B ($n = 8$), SCANet-M ($n = 16$), and SCANet-L ($n = 32$). n is the number of SCFBlock. Besides, we set the downsampling scale factor of LR-HSI to 4× during the experiment.

Overall (Table I), the results in the horizontal error direction are generally better than those in the diagonal direction. The accuracy of PSNR and SSIM gradually increases with the number of SCFBlocks, which proves beneficial to the extraction of image abstract features. When the number of SCFBlocks increases from 2 to 8, the PSNR in the horizontal error direction improves by 1.29 dB, and the SSIM improves by 0.009 dB; when the number of SCFBlocks increases from 8 to 32, the PSNR in the horizontal error direction improves by only 0.17 dB, and the SSIM is almost unchanged. The trend of the diagonal error direction is consistent with this result. SAM can respond to the spectral similarity between images. The value of SAM gradually decreases with the number of SCFBlocks, indicating that the increase in the number of blocks is beneficial to the retention of spectral

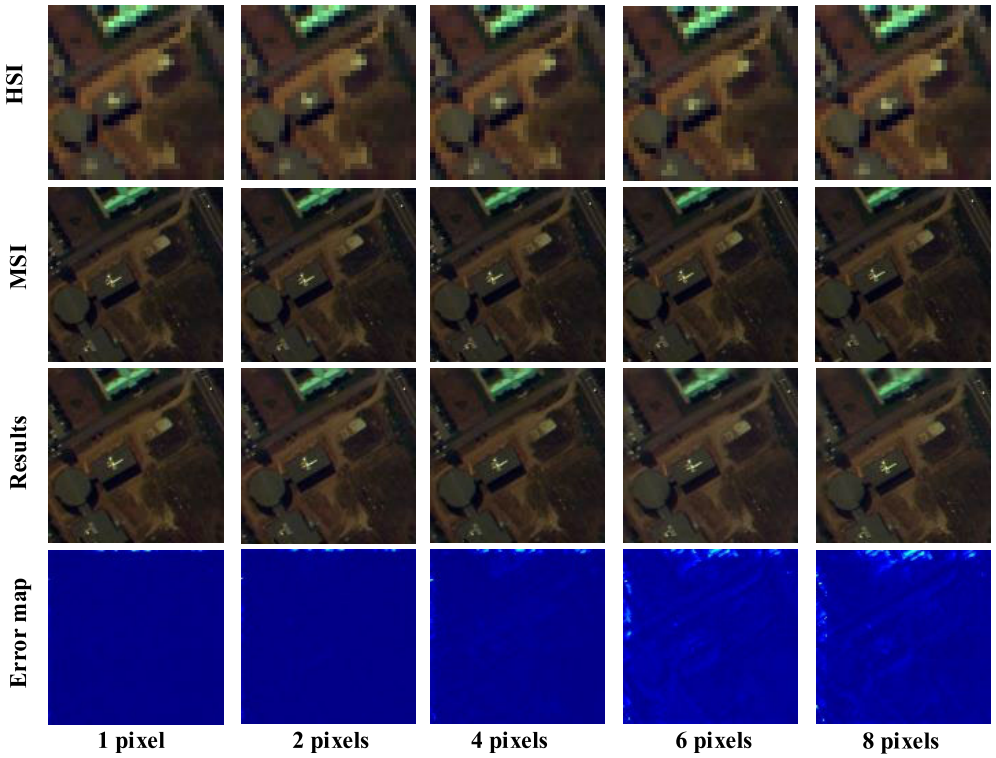

Fig. 7. Fusion results of SCANet in the diagonal registration error direction on the PaviaU dataset.

TABLE II

4× FUSION RESULTS ACHIEVED ON THE PAVIAU DATASET BY SCANET-B WITH DIFFERENT REGISTRATION ERROR DIRECTIONS. *H*, *V*, AND *D* REPRESENT THE REGISTERED ERRORS IN THE HORIZONTAL, VERTICAL, AND DIAGONAL DIRECTIONS, RESPECTIVELY. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

Rotation	Error directions	1 pixel		2 pixels		4 pixels		6 pixels		8 pixels	
		PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
0°	H	**42.53**	**0.92**	41.08	**0.92**	38.62	0.90	34.86	0.85	33.72	0.83
	V	**40.74**	**0.92**	38.67	**0.92**	36.33	0.90	34.23	0.88	32.94	0.85
	D	**40.04**	**0.92**	38.36	**0.92**	35.17	0.89	31.53	0.81	30.59	0.79
90°	H	**42.43**	**0.92**	42.01	**0.92**	39.64	0.89	37.28	0.87	35.61	0.85
	V	**40.68**	**0.92**	38.88	**0.92**	36.91	0.90	35.14	0.86	33.78	0.82
	D	**40.82**	**0.92**	38.53	**0.92**	36.10	0.88	33.81	0.83	32.14	0.78

information between images. Considering the complexity and time-running cost of the model, the SCFBlock is set to 8 by default in the subsequent experiments.

2) Registration Error Direction: We use SCANet-B to investigate the impact of registration error direction, setting the registration error range of 1–8 pixels for LR-HSI and HR-MSI.

Quantitative and qualitative results are shown in Table II and Figs. 5–7. With the increase in the registration error, the accuracy of PSNR and SSIM gradually decreases. When the registration error is one pixel, the PSNR in all three error directions (horizontal, vertical, and diagonal) is greater than 40 dB. In remote sensing applications, the registration accuracy of images is generally required to be less than 0.5 pixels, and this result shows the excellent fusion performance of SCANet-B. When the registration error reaches eight pixels, it means that the spatial positions of LR-HSI and HR-MSI differ by $1.3 \times 8 = 10.4$ m, and at this time, the feature types at the same point change, making image fusion extremely difficult. However, the PSNR in the horizontal error direction is 33.72 dB, and SSIM is 0.83 dB; there are more consistent results in the vertical error direction; the PSNR in the diagonal error direction is 30.59 dB, and the SSIM is 0.79 dB. SCANet-B can also obtain satisfactory results.

It is worth noting that horizontal and vertical registration errors theoretically have the same effect on the fusion performance, while diagonal registration errors result in worse fusion performance. However, the PSNR of horizontal registration errors is consistently better than that of vertical and diagonal registration errors (Table II, rotation = 0°). Two main aspects affect the registration error direction fusion performance: the SCANet structure itself and the CCFSA. SCANet is designed to mimic the structure of a binocular stereo camera. Its imaging process is similar to the horizontal and vertical parallaxes that occur when the left and right eyes of a person observe the same object separately. However, the perception of vertical parallax by the human eye is weaker than that of the horizontal direction [63]. Therefore, it can be seen from Table II that the PSNR of the horizontal registration error direction is always the highest, as the error increases. Moreover, CCFSA is learning cross-complementary attention, collecting contextual information in horizontal and vertical directions, and fusing

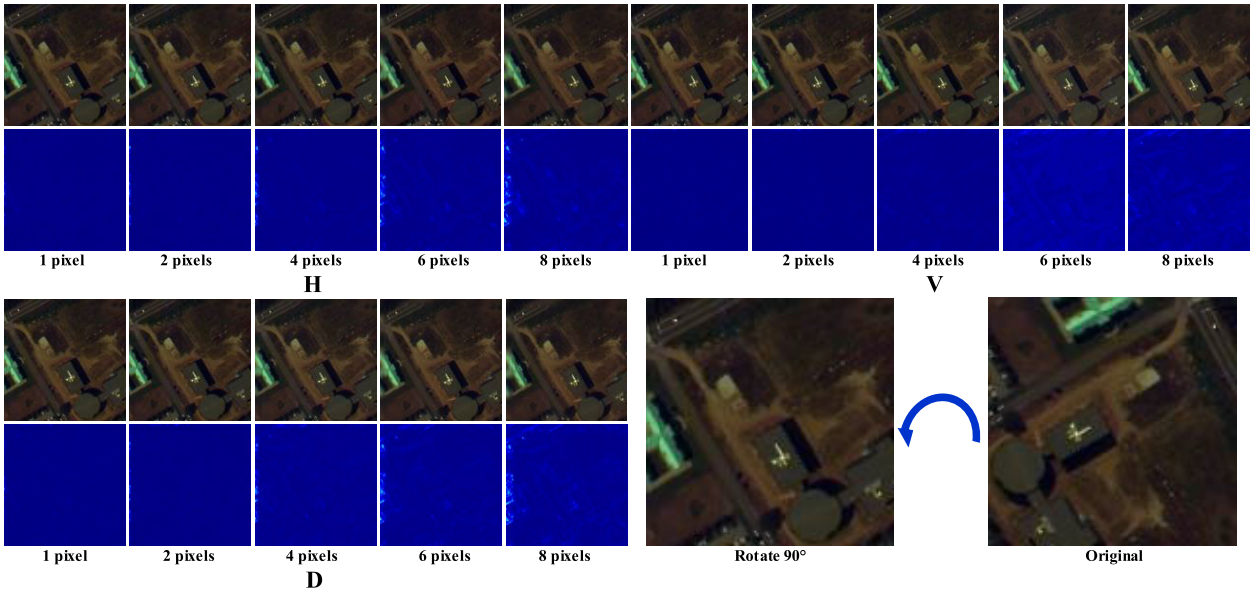

Fig. 8. Fusion results of rotating the PaviaU dataset 90° counterclockwise in the H, V, and D registration error directions, respectively.

TABLE III

FUSION RESULTS ACHIEVED ON THE PAVIAU DATASET BY SCANET-B WITH DIFFERENT ACTIVATION FUNCTIONS. H AND D REPRESENT THE REGISTERED ERRORS IN HORIZONTAL AND DIAGONAL DIRECTIONS. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

	PSNR↑		RMSE↓		ERGAS↓		SAM↓		SSIM↑		Training
	H	D	H	D	H	D	H	D	H	D	Time(Ms)
$\mathbf{X_1} \otimes \mathbf{X_2}$	**41.08**	**38.36**	**2.04**	**3.05**	1.42	**2.04**	**1.91**	**2.00**	**0.92**	**0.92**	1786.23
$\mathbf{X_1} \oplus \mathbf{X_2}$	40.48	37.21	2.32	3.38	1.56	2.24	2.05	2.23	**0.92**	0.91	**1696.97**
GeLU	41.07	37.98	2.09	3.10	1.44	2.07	1.94	2.02	**0.92**	**0.92**	1980.37
ReLU	40.58	38.11	2.17	3.25	**1.36**	2.12	2.17	2.06	**0.92**	**0.92**	1895.69

unregistered images using cross-visual information from multiple directions. It compensates to some extent the deficiency of SCANet in the vertical error direction. As the registration error increases (Table II, rotation = 0°), the PSNR in the vertical direction gradually approaches that in the horizontal direction (ΔPSNR of 2.41 (two pixels), 2.29 (four pixels), 0.63 (six pixels), and 0.78 (eight pixels) for the horizontal and vertical error directions, respectively). This analysis leads to the conclusion that the role of CCFSA, especially the ability to fuse the interaction information in the vertical error direction, gradually strengthens, as the registration error increases.

To verify the above inference, we reinput the original image into SCANet by rotating it 90° counterclockwise. Observing the fusion results in the H, V, and D error directions in the range of 1–8 pixel errors (Fig. 8). Table II (rotation = 90°) shows that as the registration error increases, the PSNR in the vertical direction gradually approaches that in the horizontal direction (ΔPSNR of 3.13 (two pixels), 2.73 (four pixels), 2.14 (six pixels), and 1.83 (eight pixels) for the horizontal and vertical error directions, respectively). The fusion results with the original images maintain a consistent trend of change.

We show the visual results of PaviaU in different registration error directions (from Figs. 5 to 8). These results suggest that our SCANet-B reconstructs satisfactory fused images with rich details and sharp edges. This demonstrates the powerful ability of SCANet-B to fully learn complementary information between images.

TABLE IV

FOUR PIXELS REGISTRATION ERROR FUSION RESULTS ACHIEVED ON THE PAVIAU DATASET BY SCANET-B WITH DIFFERENT DOWNSAMPLING SCALE FACTORS. H AND D REPRESENT THE REGISTERED ERRORS IN HORIZONTAL AND DIAGONAL DIRECTIONS. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

	PSNR↑		RMSE↓		ERGAS↓		SAM↓		SSIM↑	
	H	D	H	D	H	D	H	D	H	D
2×	**39.03**	**36.43**	**2.74**	**3.70**	**1.78**	**2.31**	2.23	**2.49**	**0.91**	**0.89**
4×	38.62	35.17	2.87	4.28	1.83	2.76	2.25	2.77	0.90	**0.89**
8×	38.20	33.74	3.02	5.05	1.88	3.09	**2.20**	3.02	0.90	0.87

3) Activation Function: We designed the SCFBlock using a gate unit instead of the nonlinear activation function in Restormer. Therefore, we compare the fusion effect of four different activation functions to verify the simplicity and effectiveness of the gate unit. Besides, we set the registration error of LR-HSI and HR-MSI to two pixels in the experiments. ⊕ in Table III represents the element-by-element summation.

Although GeLU and ReLU are the most commonly used activation functions, the PSNR of $\mathbf{X_1} \otimes \mathbf{X_2}$ improved by 0.01 dB over GeLU and 0.5 dB over ReLU (Table III). The

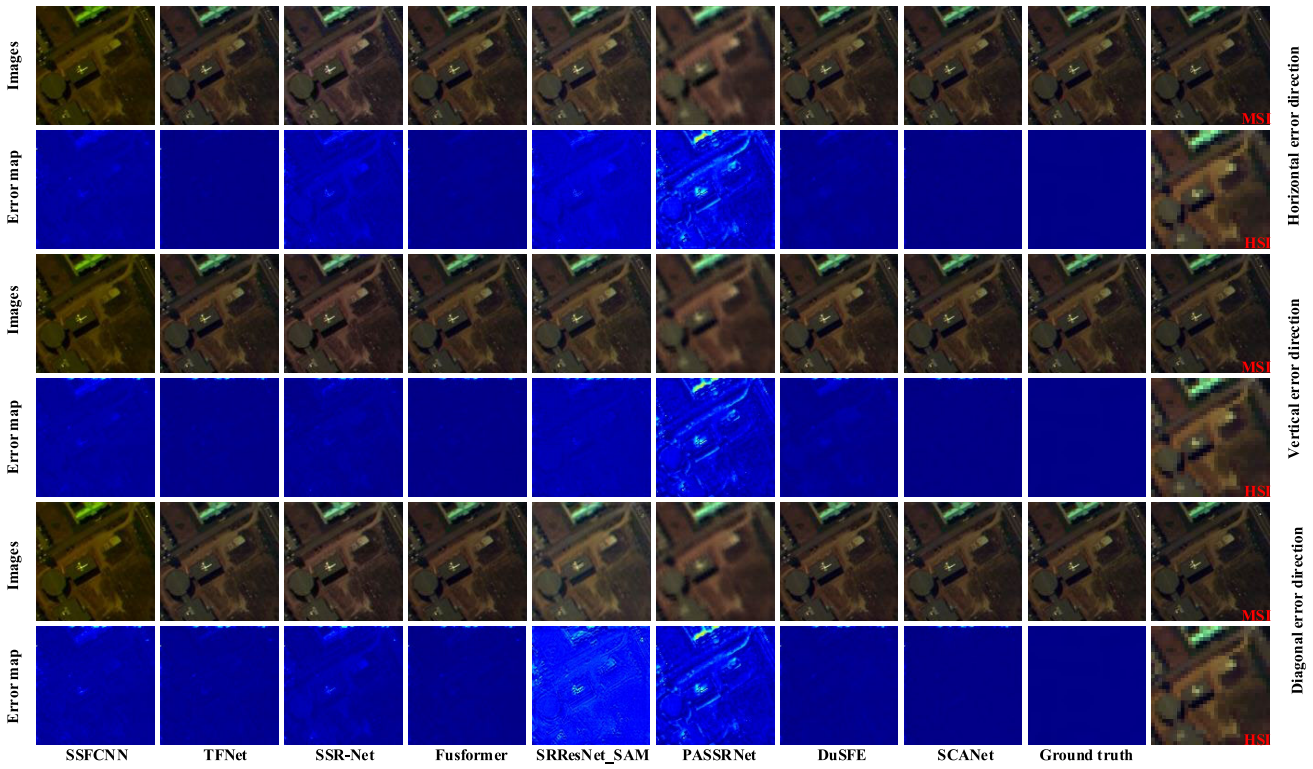

Fig. 9. Fusion results of different deep learning methods on the PaviaU dataset.

TABLE V
FUSION RESULTS ACHIEVED ON THE PAVIAU DATASET BY SCANET WITH DIFFERENT TYPES OF CROSS-FUSION MODULES. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

Modules	PSNR↑	RMSE↓	ERGAS↓	SAM↓	SSIM↑	Model size(M)
SCANet+DuAtt [65]	40.99	2.19	**1.49**	2.02	**0.92**	2.53
SCANet+CoordAtt [66]	39.76	2.52	1.67	2.30	0.90	**2.18**
SCANet+SAM [67]	33.09	3.74	2.51	3.03	0.88	3.62
SCANet+CCFSA	**41.08**	**2.17**	**1.49**	**2.01**	**0.92**	2.28

training time can reflect the running cost of the model. With the number of the training set to 2000 iterations, the training time of $\mathbf{X_1} \otimes \mathbf{X_2}$ is slightly less than that of GeLU and ReLU (194.14 Ms less than GeLU and 109.46 Ms less than ReLU). $\mathbf{X_1} \otimes \mathbf{X_2}$ has a slight advantage in operational efficiency. $\mathbf{X_1} \oplus \mathbf{X_2}$ has a simple structure and runs faster, but its fusion effect is slightly worse. The above analysis concludes that $\mathbf{X_1} \otimes \mathbf{X_2}$ has a simple and efficient structure while satisfying a good fusion effect.

4) Downsampling Scale Factor: We investigate the fusion effect of LR-HSI with HR-MSI at different downsampling ratios by comparing the accuracy of PSNR, RMSE, ERGAS, SAM, and SSIM, under the condition that the registered error is four pixels.

As the scale of downsampling increases, the spatial resolution of LR-HSI decreases, the spectral information of the objects in the pixel is complicated, the boundary of the feature is blurred, and the fusion with HR-MSI becomes gradually more difficult. Comparing the PSNR in the horizontal error direction in Table IV, 8× decreases by 0.83 dB compared with 2×; while in the diagonal error direction, 8× decreases by 2.69 dB compared with 2×. In terms of spatial fusion, RMSE and SSIM can reflect the spatial details and structural information of the fused images. The RMSE increases from 2.74 dB (2×) to 3.02 dB (8×) in the horizontal error direction and from 3.70 dB (2×) to 5.05 dB (8×) in the diagonal error direction. SSIM remains almost unchanged in the horizontal error direction, while it decreases by 0.02 dB in the diagonal error direction. The ERGAS for 2×, 4×, and 8× in the horizontal error direction is 1.78, 1.83, and 1.88 dB, respectively, and the ΔERGAS is 0.1 dB, while the ΔERGAS in the diagonal error direction is 0.78 dB. The SAM for 2×, 4×, and 8× in the horizontal error direction is 2.23, 2.25, and 2.20 dB, respectively, and the ΔSAM is 0.03 dB, while the ΔSAM in the diagonal error direction is 0.53 dB. The above analysis leads to the conclusion that the downsampling scale of LR-HSI is more difficult to reconstruct the spatial and spectral information of the image in the diagonal error direction compared with the horizontal error direction.

5) Types of Cross-Fusion Modules: We compare the results of SCFBlock with different cross-fusion modules to investigate the potential influence introduced by different design choices. DuAtt [64], CoordAtt [65], SAM [66], and CCFSA were mainly selected for the experiments. Besides, we set LR-HSI and HR-MSI as the horizontal registration error direction with two pixels.

As shown in Table V, all models with SCFBlock achieved good results. Specifically, SCANet + CCFSA can achieve 0.09–7.99-dB improvement in PSNR, 0.02–1.57-dB

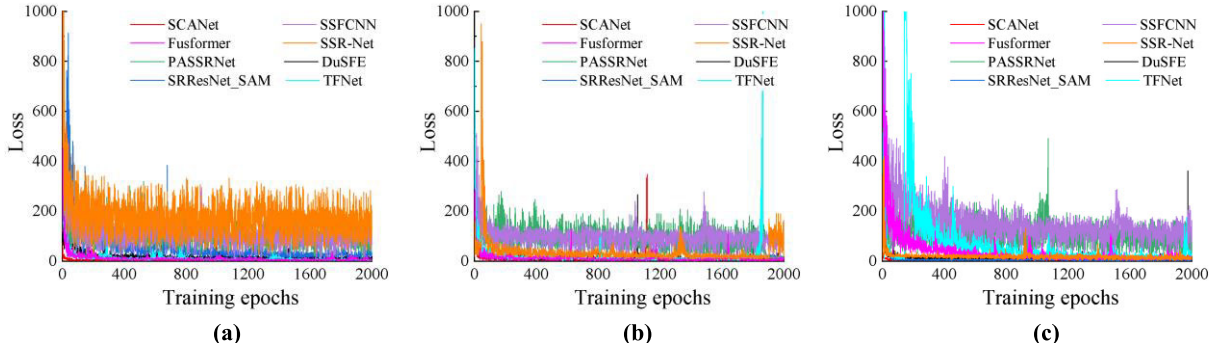

Fig. 10. Loss function of different deep learning methods. (a)–(c) Loss functions for horizontal, vertical, and diagonal registration error directions.

TABLE VI

FUSION RESULTS OF DIFFERENT DEEP LEARNING METHODS ON PAVIAU DATASETS. H, V, AND D REPRESENT THE REGISTERED ERRORS IN THE HORIZONTAL, VERTICAL, AND DIAGONAL DIRECTIONS, RESPECTIVELY. THE BEST DATA ARE MARKED IN BOLD. THE ARROW ATTACHED TO THE METRICS POINTS TO BETTER PERFORMANCE

Methods	H			V			D			Model Size(M)	Running Times(Ms)
	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓	PSNR↑	SSIM↑	SAM↓		
SSFCNN [25]	27.68	0.770	11.181	27.58	0.783	11.125	26.586	0.703	12.192	1.76	75.85
TFNet [27]	39.85	0.907	2.321	37.47	0.901	2.571	35.42	0.867	3.057	9.36	105.23
SSR-Net [28]	34.54	0.819	3.479	34.87	0.852	3.096	33.11	0.795	5.246	**1.1**	**51.86**
Fusformer [30]	39.94	0.90	2.356	37.944	0.904	2.464	36.30	0.881	2.838	8.89	98.25
SRResNet_SAM [67]	35.32	0.830	4.182	35.30	0.855	3.844	34.94	0.892	4.318	21.09	3216.99
PASSRNet [68]	25.71	0.358	6.446	25.79	0.374	6.177	25.75	0.376	6.219	9.91	85.77
DuSFE [69]	37.67	0.885	2.850	37.19	0.894	2.689	36.42	0.894	2.827	77.46	1161.67
SCANet	**41.08**	**0.932**	**2.016**	**38.67**	**0.926**	**1.977**	**38.36**	**0.920**	**2.000**	2.28	1786.23

improvement in RMSE, and 0.01–1.02-dB improvement in SAM compared with other models. It is also worth noting that the model size of SCANet + CCFSA is 0.25M lower than that of SCANet + DuAtt and 1.34M lower than that of SCANet + SAM, indicating that CCFSA makes full use of cross-complementary information between images while having a simpler structure.

C. Comparison With Different Methods

We compare SCANet with the existing deep learning fusion methods (experimented with three different registration error directions), including spectral stride fusion CNN network (SSFCNN) [25], TFNet [27], SSR-Net [28], Fusformer [30], super-resolution residual network_stereo attention module (SRResNet_SAM) [66], parallax-attention stereo super-resolution network (PASSRnet) [67], and dual-branch squeeze-fusion-excitation (DuSFE) [68]. For a fair comparison, the registration error is set to two pixels for all the methods.

1) Qualitative Evaluation: The comparative results of different fusion methods are shown in Fig. 9. In general, PASSRNet can handle large noise, but lose edge definition and detail information, and cannot effectively improve the spatial resolution, and the fusion results are the most blurred. SSFCNN, PASSRNet, and SRResNet_SAM have more obvious color distortion, which is shown in all three groups of experiments. Both SSR-Net and DuSFE can effectively improve the spatial resolution, and the texture of the features is clearer, but the performance in terms of color fidelity is slightly different. In the three groups of experiments, the color of TFNet and Fusformer fusion result is similar to the MSI, while the color of DuSFE and SSR-Net fusion result is closer to the HSI, which indicates to some extent that the spectral fidelity of DuSFE and SSR-Net is better than that of TFNet and Fusformer. The fusion result of SCANet is closest to the reference image with natural tones and shows good spectral retention and spatial resolution enhancement in all three groups of experiments.

2) Quantitative Evaluation: The comparative results of different fusion methods are shown in Table VI. In terms of spatial reconstruction, SCANet performs the best, and DuSFE and SRResNet_SAM also perform better. The SSIM of these three fusion results reached 0.8 dB in different registration error directions, and the PSNR was above 35 dB. This indicates that the spatial reconstruction ability of the method based on the stereo vision principle is little affected by the registration error. TFNet and Fusformer have excellent fusion performance in the horizontal error direction, with both PSNR reaching above 0.9 dB and SSIM of 0.907. However, their PSNR and SSIM in the vertical and diagonal error directions drop faster, making it difficult to focus on the spatial information from different directions. In terms of spectral retention, both SSR-Net and DuSFE showed good results. The spectral fidelity of SSR-Net was poor in the diagonal error direction, with SRResNet_SAM reaching 5.246 dB, but there was a large improvement in the horizontal and vertical error directions, indicating that SSR-Net was not good at processing the spectral information in the diagonal direction. In these three groups of experiments, the SAMs of SSFCNN and PASSRNet reach 11 and 6 dB or more, respectively, with low spectral retention. The results clearly show the effectiveness of the SCANet.

3) Runtime Efficiency: We also report the runtime (evaluated on NVIDIA NVS 310 GPUs with 128 × 128 inputs, Table VI) and the variation of the loss function (Fig. 10) to compare the computational complexity among SSFCNN,

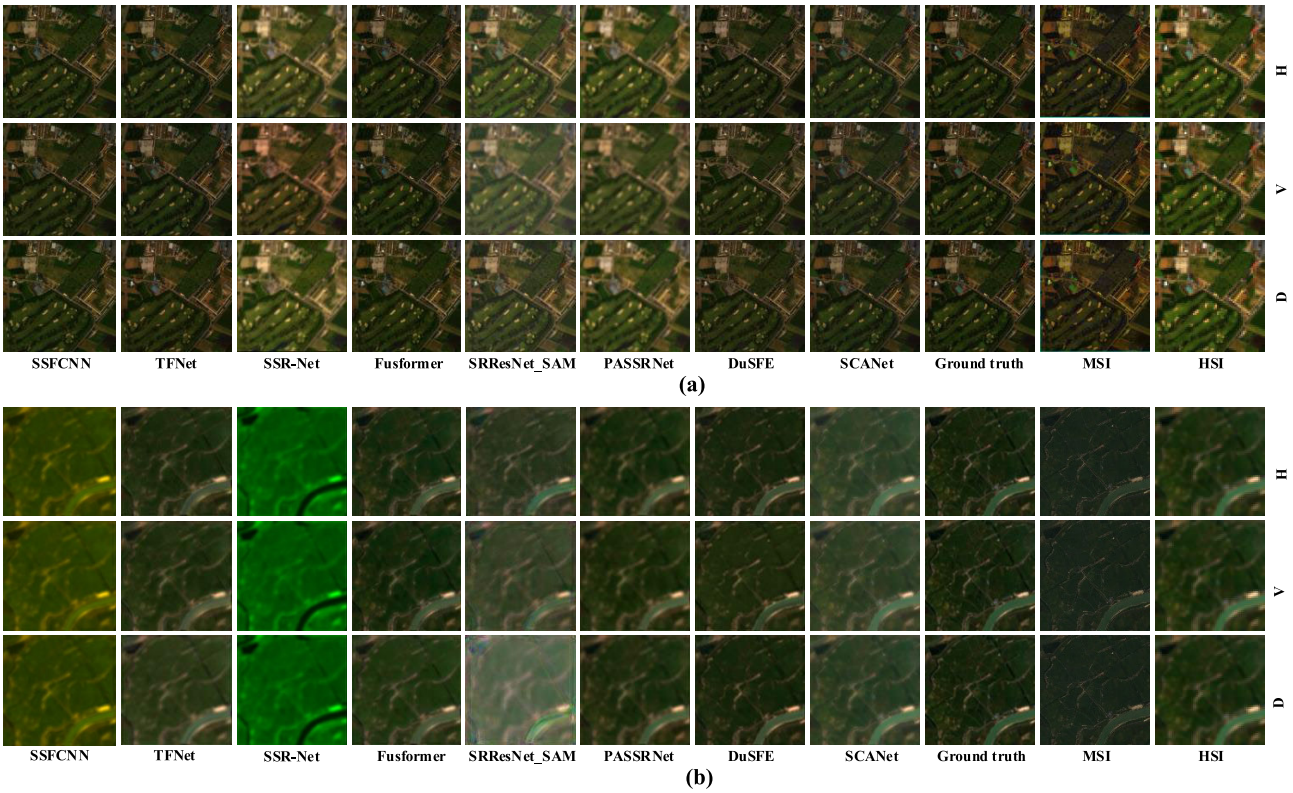

Fig. 11. Fusion results of different deep learning methods on (a) Chikusei and (b) PYLake datasets. H, V, and D represent the registered errors in the horizontal, vertical, and diagonal directions, respectively.

TFNet, SSR-Net, Fusformer, SRResNet_SAM, PASSRNet, DuSFE, and SCANet. SCANet has a PSNR of 38.36–41.08 dB for the experiments in the three registration error directions, with a 34.59% improvement in runtime over SRResNet_SAM (Table VI), and the loss function converges at a faster rate (Fig. 10). Considering the fusion quality again, the proposed SCANet has the overall advantage in the fusion of LR-HSI and HR-MSI.

D. Application of Real Remote Sensing Images

In practical applications, there are large gaps between multisource remote sensing images, including data source, acquisition time, spatial resolution, spectral number, and range. Therefore, we used Chikusei and PYLake datasets to verify the generalizability of SCANet. We likewise compared the results of SSFCNN [25], TFNet [27], SSR-Net [28], Fusformer [30], SRResNet_SAM [66], PASSRNet [67], and DuSFE [68] in three registration error directions. The registration error is set to two pixels.

1) Chikusei Dataset: The downsampling ratio of LR-HSI is set to 4, i.e., we fused HR-MSI with a spatial resolution of 2.5 m and LR-HSI with a spatial resolution of 10 m. The size of the Chikusei dataset is 800 × 800 pixels, the subregion of 256 × 256 pixels in the center is cropped as the test image, and the rest region is used for training. Therefore, more training samples are available compared with the PYLake dataset. It can be visually observed that the SCANet greatly improves the edge and feature texture information in the image, which is superior to other fusion results (Fig. 11).

2) PYLake Dataset: The spatial resolution of HR-MSI in the PYLake dataset is 30 m, and the downsampling ratio of LR-HSI is set to 4, i.e., we fused HR-MSI with a spatial resolution of 30 m and LR-HSI with a spatial resolution of 120 m. The PYLake dataset has complex feature types and fragmented spatial distribution. When the registration error is two pixels, the feature types at the same point in the image may change, thus inevitably causing spatial and spectral distortions during the fusion process. Fig. 11 shows that the fusion results of all eight methods are significantly degraded compared with the PaviaU and Chikusei datasets. SSFCNN and SSR-Net produce severe spectral distortions. The loss function of SSFCNN and SSR-Net can very well display the spatial and spectral edge information. However, due to the inaccuracy of HSI in performing upsampling operations, high-frequency edge textures are lost compared with MSI, and direct cross-channel fusion produces structural distortion. The texture in the PASSRNet and SRResNet_SAM fusion results is blurred. The color of the TFNet, Fusformer, and DuSFE fusion results is close to MSI, while the SCANet fusion results are close to HSI. Both DuSFE and SCANet showed good spatial resolution enhancement ability in the three experiments.

The fusion effect of SCANet varies with specific variations when processing different images under different acquisition conditions. However, it can be concluded from the experimental results that the SCANet achieves desirable results in terms of robustness and fusion performance in the fusion of unregistered LR-HSI and HR-MSI.

VI. Conclusion

Considering the actual situation of remote sensing applications, we propose an SCANet for the first time based on the principle of stereo vision to solve the problem of image detail information loss due to the misregistration of LR-HSI and

HR-MSI. To avoid model complexity and improve operational efficiency, we design a simple and stackable SCFBlock for abstract feature extraction of the image. Moreover, the CCFSA performs bidirectional cross-vision interaction by combining the complementary features generated by the SCFBlock and fuses the interaction information with the input image features to achieve full utilization of horizontal and vertical contextual information of LR-HSI and HR-MSI. The fusion effects on different datasets and different settings, such as registration error directions, model sizes, and module effects, are evaluated in the experiments. The experimental results show that for the fusion of unregistered LR-HSI and HR-MSI, SCANet surpasses the current deep learning models, achieving state-of-the-art performance. In the future, we will design efficient and reliable model structures for more complex cases (e.g., squeezed, stretched, and so on) to solve the problem of HSI and MSI fusion due to the lack of strict registration.

REFERENCES

[1] S.-E. Qian, "Hyperspectral satellites, evolution, and development history," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 7032–7056, 2021.

[2] R. A. Borsoi, C. Prévost, K. Usevich, D. Brie, J. C. M. Bermudez, and C. Richard, "Coupled tensor decomposition for hyperspectral and multispectral image fusion with inter-image variability," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 702–717, Apr. 2021.

[3] R. Dian, S. Li, B. Sun, and A. Guo, "Recent advances and new guidelines on hyperspectral and multispectral image fusion," *Inf. Fusion*, vol. 69, pp. 40–51, May 2021.

[4] P. Xiang, J. Song, H. Qin, W. Tan, H. Li, and H. Zhou, "Visual attention and background subtraction with adaptive weight for hyperspectral anomaly detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 2270–2283, 2021.

[5] L. Tang, Z. Li, W. Wang, B. Zhao, Y. Pan, and Y. Tian, "An efficient and robust framework for hyperspectral anomaly detection," *Remote Sens.*, vol. 13, no. 21, p. 4247, Oct. 2021.

[6] X. Fu, S. Jia, L. Zhuang, M. Xu, J. Zhou, and Q. Li, "Hyperspectral anomaly detection via deep plug-and-play denoising CNN regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 11, pp. 9553–9568, Nov. 2021.

[7] W. Yu, M. Zhang, and Y. Shen, "Spatial revising variational autoencoder-based feature extraction method for hyperspectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 2, pp. 1410–1423, Feb. 2021.

[8] A. Essa, P. Sidike, and V. Asari, "Volumetric directional pattern for spatial feature extraction in hyperspectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 7, pp. 1056–1060, Jul. 2017.

[9] C. Zou and Y. Xia, "Bayesian dictionary learning for hyperspectral image super resolution in mixed Poisson–Gaussian noise," *Signal Process., Image Commun.*, vol. 60, pp. 29–41, Feb. 2018.

[10] A. Sellami, "Semantic interpretation of hyperspectral imagery based on adaptive dimensionality reduction," Ph.D. dissertation, Université de Lille, Lille, France, Dec. 2017. [Online]. Available: https://www.researchgate.net/publication/323265192_Semantic_Interpretation_of_Hyperspectral_Imagery_based_on_Adaptive_Dimensionality_Reduction

[11] W. Zhao, S. Li, A. Li, B. Zhang, and J. Chen, "Deep fusion of hyperspectral images and multi-source remote sensing data for classification with convolutional neural network," *Nat. Remote Sens. Bull.*, vol. 25, no. 7, pp. 1489–1502, 2021.

[12] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.

[13] S. Zhong, Y. Zhang, Y. Chen, and D. Wu, "Combining component substitution and multiresolution analysis: A novel generalized BDSD pansharpening algorithm," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2867–2875, Jun. 2017.

[14] W. Dong, Y. Yang, J. Qu, S. Xiao, and Q. Du, "Hyperspectral pansharpening via local intensity component and local injection gain estimation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[15] M. Maneshi, H. Ghassemian, G. Khademi, and M. Imani, "A retina-inspired multiresolution analysis framework for pansharpening," in *Proc. Int. Conf. Mach. Vis. Image Process. (MVIP)*, Feb. 2020, pp. 1–5.

[16] G. Licciardi, G. Vivone, M. D. Mura, R. Restaino, and J. Chanussot, "Multi-resolution analysis techniques and nonlinear PCA for hybrid pansharpening applications," *Multidimensional Syst. Signal Process.*, vol. 27, no. 4, pp. 807–830, Oct. 2016.

[17] K. Sawada, K. Hashimoto, Y. Nankaku, and K. Tokuda, "A Bayesian framework for image recognition based on hidden Markov eigen-image models," *IEEJ Trans. Electr. Electron. Eng.*, vol. 13, no. 9, pp. 1335–1347, Sep. 2018.

[18] J. Xiao, R. Zhao, and K.-M. Lam, "Bayesian sparse hierarchical model for image denoising," *Signal Process., Image Commun.*, vol. 96, Aug. 2021, Art. no. 116299.

[19] Q. Li, H. Li, Z. Lu, Q. Lu, and W. Li, "Denoising of hyperspectral images employing two-phase matrix decomposition," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 9, pp. 3742–3754, Sep. 2014.

[20] H. Chang, T. Wang, A. Li, and H. Fang, "Local hyperspectral anomaly detection method based on low-rank and sparse matrix decomposition," *J. Appl. Remote Sens.*, vol. 13, no. 2, p. 1, Jun. 2019.

[21] X. Fu, S. Jia, M. Xu, J. Zhou, and Q. Li, "Fusion of hyperspectral and multispectral images accounting for localized inter-image changes," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5517218.

[22] K. Ksiek, P. Gomb, M. Romaszewski, M. Cholewa, and B. Grabowski, "Stable training of autoencoders for hyperspectral unmixing," 2021. [Online]. Available: https://www.researchgate.net/publication/354928469_Stable_training_of_autoencoders_for_hyperspectral_unmixing

[23] W. Wang, W. Zeng, Y. Huang, X. Ding, and J. Paisley, "Deep blind hyperspectral image fusion," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4149–4158.

[24] Z. Wang, B. Chen, R. Lu, H. Zhang, H. Liu, and P. K. Varshney, "FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 7565–7577, 2020.

[25] X.-H. Han, B. Shi, and Y. Zheng, "SSF-CNN: Spatial and spectral fusion with CNN for hyperspectral image super-resolution," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 2506–2510.

[26] Q. Yang, Y. Xu, Z. Wu, and Z. Wei, "Hyperspectral and multispectral image fusion based on deep attention network," in *Proc. 10th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2019, pp. 1–5.

[27] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, Mar. 2020.

[28] X. Zhang, W. Huang, Q. Wang, and X. Li, "SSR-Net: Spatial–spectral reconstruction network for hyperspectral and multispectral image fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5953–5965, Jul. 2021.

[29] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[30] J.-F. Hu, T.-Z. Huang, L.-J. Deng, H.-X. Dou, D. Hong, and G. Vivone, "Fusformer: A transformer-based fusion network for hyperspectral image super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[31] Y. Qu, H. Qi, C. Kwan, N. Yokoya, and J. Chanussot, "Unsupervised and unregistered hyperspectral image super-resolution with mutual Dirichlet-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5507018.

[32] D. Mistry and A. Banerjee, "Review: Image registration," *Int. J. Graph. Image Process.*, vol. 2, pp. 18–22, Feb. 2012.

[33] X. Zhang, C. Leng, Y. Hong, Z. Pei, I. Cheng, and A. Basu, "Multimodal remote sensing image registration methods and advancements: A survey," *Remote Sens.*, vol. 13, no. 24, p. 5128, Dec. 2021.

[34] Y. Tong, H. Li, J. Chen, M. Zhao, and L. Liu, "Dual-band stereo vision based on heterogeneous sensor networks," *Signal Process.*, vol. 126, pp. 87–95, Sep. 2016.

[35] Y. Yang, M. Gao, J. Zhang, Z. Zha, and Z. Wang, "Depth map super-resolution using stereo-vision-assisted model," *Neurocomputing*, vol. 149, pp. 1396–1406, Feb. 2015.

[36] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5718–5729.

[37] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[38] W. Wang, L. Jiao, and S. Yang, "Novel adaptive component-substitution-based pan-sharpening using particle swarm optimization," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 781–785, Apr. 2015.

[39] X. Huang and L. Zhang, "A multilevel decision fusion approach for urban mapping using very high-resolution multi/hyperspectral imagery," *Int. J. Remote Sens.*, vol. 33, no. 11, pp. 3354–3372, Jun. 2012.
[40] X. Hu, Y. Shi, W. Li, and R. Tao, "Improved multiresolution analysis method for hyperspectral pansharpening," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 2778–2781.
[41] L. Loncan et al., "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.
[42] K. Ren, W. Sun, X. Meng, G. Yang, J. Peng, and J. Huang, "A locally optimized model for hyperspectral and multispectral images fusion," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5519015.
[43] K. Wang, Y. Wang, X.-L. Zhao, D. Meng, J. Cheung, and Z. Xu, "Hyperspectral and multisectral image fusion via nonlocal low-rank tensor decomposition and spectral unmixing," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 550–562, Jan. 2021.
[44] J. Mou, W. Gao, and Z. Song, "Image fusion based on non-negative matrix factorization and infrared feature extraction," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, vol. 2, Dec. 2013, pp. 1046–1050.
[45] N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled non-negative matrix factorization (CNMF) for hyperspectral and multispectral data fusion: Application to pasture classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 1779–1782.
[46] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.
[47] L. He, J. Zhu, J. Li, A. Plaza, J. Chanussot, and B. Li, "HyperPNN: Hyperspectral pansharpening via spectrally predictive convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 3092–3100, Aug. 2019.
[48] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.
[49] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.
[50] S. Jiamin and S. Huihui, "Hyperspectral and multispectral image fusion based on discrete wavelet transform and generative adversarial networks," *Radio Eng.*, vol. 51, no. 12, pp. 1434–1441, 2021.
[51] H. Zhou, Q. Liu, and Y. Wang, "PanFormer: A transformer based model for pan-sharpening," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.
[52] V. Vs, J. M. Jose Valanarasu, P. Oza, and V. M. Patel, "Image fusion transformer," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 3566–3570.
[53] L. Qu, S. Liu, M. Wang, and Z. Song, "TransMEF: A transformer-based multi-exposure image fusion framework using self-supervised multi-task learning," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 2126–2134.
[54] X. Wang, X. Wang, R. Song, X. Zhao, and K. Zhao, "MCT-Net: Multi-hierarchical cross transformer for hyperspectral and multispectral image fusion," *Knowl.-Based Syst.*, vol. 264, Mar. 2023, Art. no. 110362.
[55] A. Guo, R. Dian, and S. Li, "A deep framework for hyperspectral image fusion between different satellites," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 7939–7954, Jul. 2023.
[56] Y. Zhou, A. Rangarajan, and P. D. Gader, "An integrated approach to registration and fusion of hyperspectral and multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3020–3033, May 2020.
[57] J. Nie, L. Zhang, W. Wei, C. Ding, and Y. Zhang, "Unsupervised deep hyperspectral super-resolution with unregistered images," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.
[58] Y. Fu, Y. Zheng, L. Zhang, Y. Zheng, and H. Huang, "Simultaneous hyperspectral image super-resolution and geometric alignment with a hybrid camera system," *Neurocomputing*, vol. 384, pp. 282–294, Apr. 2020.
[59] K. Zheng, L. Gao, D. Hong, B. Zhang, and J. Chanussot, "NonRegSRNet: A nonrigid registration hyperspectral super-resolution network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5520216.
[60] P. Firoozfam, "Multicamera imaging for three-dimensional mapping and positioning: stereo and panoramic conical views," Jan. 2004. [Online]. Available: https://www.researchgate.net/publication/254696597_Multicamera_imaging_for_three-dimensional_mapping_and_positioning_Stereo_and_panoramic_conical_views
[61] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.
[62] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[63] J. Hu, S. Li, Y. Chang, and C. Hou, "Comfortable disparity range of stereo image based on salient region," *Acta Optica Sinica*, vol. 38, no. 8, 2018, Art. no. 0811001.
[64] S. Wang, D. Liang, J. Song, Y. Li, and W. Wu, "DABERT: Dual attention enhanced BERT for semantic matching," in *Proc. 29th Int. Conf. Comput. Linguistics*, Oct. 2022, pp. 1645–1654.
[65] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.
[66] X. Ying, Y. Wang, L. Wang, W. Sheng, W. An, and Y. Guo, "A stereo attention module for stereo image super-resolution," *IEEE Signal Process. Lett.*, vol. 27, pp. 496–500, 2020.
[67] L. Wang et al., "Learning parallax attention for stereo image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12242–12251.
[68] X. Chen et al., "Dual-branch squeeze-fusion-excitation module for cross-modality registration of cardiac SPECT and CT," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds. Cham, Switzerland: Springer, 2022, pp. 46–55.

Yujuan Guo received the M.E. degree in geographic information systems from the Xi'an University of Science and Technology, Xi'an, China, in 2015, and the Ph.D. degree in geographic information systems from the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China, in 2021.

She is currently a Post-Doctoral Researcher with Shenzhen University, Shenzhen, China. Her research interests include hyperspectral image fusion, image segmentation, and deep learning.

Xiyou Fu (Member, IEEE) received the bachelor's degree from Wuhan University, Wuhan, China, in 2012, and the M.S. and Ph.D. degrees from the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2015 and 2019, respectively.

He is currently an Associate Researcher Fellow with Shenzhen University, Shenzhen, China. His research interests include hyperspectral image restoration, anomaly detection, and super-resolution.

Meng Xu (Member, IEEE) received the B.S. and M.E. degrees in electrical engineering from the Ocean University of China, Qingdao, China, in 2011 and 2013, respectively, and the Ph.D. degree from the University of New South Wales, Canberra, ACT, Australia, in 2017.

She is currently an Assistant Professor with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China. Her research interests include cloud removal and remote sensing image processing.

Sen Jia (Senior Member, IEEE) received the B.S. and Ph.D. degrees from the College of Computer Science, Zhejiang University, Hangzhou, China, in 2002 and 2007, respectively.

Since 2008, he has been with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, where he is currently a Full Professor. His research interests include hyperspectral image processing, signal and image processing, and machine learning.