Cross-Domain Association Mining Based Generative Adversarial Network for Pansharpening

Lijun He, Wanyue Zhang, Jiankang Shi, and Fan Li, *Senior Member, IEEE*

***Abstract*—Multispectral (MS) pansharpening can improve the spatial resolution of MS images, which plays an increasingly important role in agriculture and environmental monitoring. Existing neural network-based methods tend to focus on global features of images, without considering the inherent relationships between similar substances in MS images. However, there is a high probability that different substances at the junction mix with each other, which leads to spectral distortion in the final pansharpened image. In this article, we propose a cross-domain association mining-based generative adversarial network for pansharpening, which consists of a spectral fidelity generator and dual discriminators. In our spectral fidelity generator, the cross-region similarity attention module is designed to establish dependencies between similar substances at different positions in the image, thereby leveraging the similar spectral features to generate pansharpened images with better spectral preservation. To mine the potential relationship between the MS image domain and the panchromatic image domain, we pretrain a spatial information extraction network. The network is then transferred to the dual-discriminator architecture to obtain the spatial information of the pansharpened images more accurately and prevent the loss of spatial details. The experimental results show that our method outperforms several state-of-the-art pansharpening methods in both quantitative and qualitative evaluations.**

***Index Terms*—Deep learning, dual discriminators, image association, multispectral (MS) pansharpening.**

I. Introduction

With the continuous launch of satellites in many countries, remote sensing images have been widely used in various fields, such as precision agriculture [1], mineral exploration [2], and ecological environment monitoring [3], [4]. And most of these applications require high-resolution multispectral (HRMS) images for higher precision and accuracy. However, due to the physical limitations, it is difficult for satellite sensors to obtain images with both high spatial resolution and high spectral resolution. They can only provide panchromatic (PAN) images and low-resolution multispectral (LRMS) images, respectively. To solve this problem, pansharpening aims to fuse the spectral information of LRMS images and the spatial information of PAN images to obtain pansharpened HRMS images.

Over the past few decades, pansharpening has attracted great attention from researchers and many pansharpening methods have emerged. Existing methods can be divided into two main categories: 1) traditional methods and 2) deep learning-based methods. Traditional methods [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] mainly rely on the detail injection model. Spatial details of PAN image are extracted in different ways, and then injected into MS image with estimated injection coefficients, which control the amount of injected details. This class of methods can obtain a tradeoff between performance and computational burden. In recent years, with the development of deep learning, there have been many works based on deep learning in the field of pansharpening. The first kinds of methods based on convolutional neural networks (CNNs), such as [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29]. These methods make improvements to the single-image super-resolution network structures or combine CNNs with traditional methods, mainly focusing on deepening the network layers to utilize the feature extraction power of CNNs. In order to obtain more realistic pansharpened images, researchers have started to explore generative adversarial networks (GANs) [30] to solve the pansharpening problem [31], [32], [33], [34], [35], [36], [37], [38]. Not only the generator is improved but also multiple discriminators are considered. Compared with CNNs, GANs can better fit the distribution of input data due to the adversarial learning between the generator and the discriminator.

Although deep learning-based methods can achieve remarkable performance by exploiting the powerful representation learning capability of neural networks, they mainly focus on global information and ignore the characteristics of MS images. In fact, cross-region similarity exists in MS images. The MS image contains many repetitive regions with similar substances, and the spectral features of these similar substances are also very similar. Therefore, this relationship between similar substances can be exploited to improve the pansharpening quality. As shown in the example of Fig. 1, M and N represent two different substances located at the junction. When pansharpening is performed on the junction, the two different substances will interact with each other, resulting in spectral distortion. However, by

Manuscript received 30 May 2022; revised 27 July 2022 and 22 August 2022; accepted 1 September 2022. Date of publication 7 September 2022; date of current version 16 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U1903213, in part by the Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology under Grant 2021SYS-04, and in part by the Natural Science Foundation of Sichuan Province 2022NSFSC0966. *(Corresponding atuhor: Lijun He.)*

Lijun He, Wanyue Zhang, and Fan Li are with the Shaanxi Key Laboratory of Deep Space Exploration Intelligent Information Technology, School of Information and Communications Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: lijunhe@mail.xjtu.edu.cn; zhangwanyue@stu.xjtu.edu.cn; lifan@mail.xjtu.edu.cn).

Jiankang Shi is with the School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: jiankang@stu.xjtu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3204824

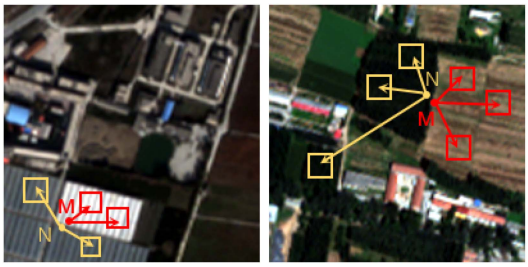

Fig. 1. Example of multispectral images to illustrate its characteristics. Each subfigure shows two representative similar regions, represented by boxes of the same color. The different substances at the junction (denoted by M and N, respectively) utilize the features of similar regions to generate similar objects, thus reducing the adverse effect of mixing different substances with each other.

utilizing spectral information from other similar substances, the mixing of substances at the junction will be avoided, thus reducing the chance of local error propagation. As shown by the partial arrows in the Fig. 1, M and N, respectively, focus on substances that are similar to them and take advantage of their similar spectral features, thereby reducing the mixing of different substances. In addition, the spatial information of PAN image is partially lost, which leads to a gap between the pansharpened image and PAN image. Similar to single-image super-resolution methods, most GAN-based pansharpening methods use a single discriminator to distinguish the image as a whole, which lacks the pertinence of spatial and spectral information. A few methods based on dual discriminators can distinguish the spatial and spectral information separately, but it is inadequate to acquire and represent the spatial information of pansharpened image exactly, resulting in poor preservation of spatial information.

To address the abovementioned issues, we design a cross-domain association mining-based GAN for pansharpening. The contributions of our work can be summarized as follows.

1) *A Cross-Region Similarity Attention Module-Based Spectral Fidelity Generator:* In order to take full advantage of the promotion of homogeneous features to pansharpening, a cross-region similarity attention module is proposed to establish similarity dependencies between distant but highly similar substances. In particular, it can reduce the adverse effects of the features at the junction by other adjacent different features, avoiding severe spectral distortion.
2) *Dual-Discriminator Architecture by Transferring our Pretrained Apatial Information Extraction Network:* To accurately obtain the spatial information to be involved in the spatial discriminator, we pretrain a spatial information extraction network which can achieve the ability to obtain the spatial component information of MS image by learning the potential relationship between ground truth and PAN images. Through the adversarial learning of the spatial discriminator and the generator, the spatial information of the pansharpened image is as consistent as possible with the input PAN image. Moreover, blurring and downsampling are performed on the pansharpened image to obtain its spectral component information, which is then entered into the spectral discriminator to maintain spectral information.

The rest of this article is organized as follows. Section II summarizes the related work. Section III gives details of the proposed pansharpening method. In Section IV, the experiments and results are described. Finally, Section V concludes this article.

II. Related Work

A. Traditional Methods

Traditional methods are mainly divided into three categories, i.e., 1) component substitution (CS) methods, 2) multiresolution analysis (MRA) methods, and 3) hybrid methods. The CS methods usually separate the spatial and spectral components of MS image in the transformed domain, where the spatial component is substituted by a histogram matched version of PAN image. Then the pansharpened image is obtained by the inverse transformation back to the original domain. The widely employed transformations in CS methods mainly include intensity–hue–saturation (IHS) [5], principal component analysis (PCA) [6], Gram–Schmidt (GS) [7], and so on. The CS methods can preserve the spatial details of the PAN image, but are easily affected by the correlation between PAN image and the spatial component of MS image. The smaller the correlation between the two, the more severe the spectral distortion of the pansharpened image. The MRA methods obtain spatial details by multiresolution decomposition of PAN image, which are then injected into MS image. Some representative instances of such methods are modulation transfer function (MTF) [8], Laplace pyramid [9], wavelet transform [10], and curvelet transform [11], [12].The MRA methods mainly operate on PAN image instead of changing the original structure of MS image, so they can preserve the spectral information of MS image. However, the linear injection of detail information can lead to spatial distortion. Hybrid methods [13], [14], which are the combination of CS and MRA methods, have the advantages of these two methods, but increase the complexity and difficulty of implementation. In addition, Xiao et al. [15] propose a variational pansharpening method with context-aware details injection fidelity, which can fully explore the complicated relationship between the PAN image and the HRMS image in the gradient domain with adaptive coefficients estimation. The method is effective in extracting the main features from the two inputs to be fused.

B. CNN-Based Methods

In recent years, CNNs have received great attention in pansharpening research. Zhong et al. [16] introduced CNNs into this field for the first time, using a CNN to enhance the intensity component of MS image and then fusing it with PAN image through GS transform. Inspired by the single-image super-resolution network structure SRCNN [39], Masi et al. [17] designed a simple CNN to achieve pansharpening. Compared with the traditional methods, the performance of this method has been greatly improved. Yang et al. [18] used a residual network [40] in the pansharpening task for the first time and trained the network in the high-frequency domain to better preserve the spatial information of PAN image. Recent studies

have shown that shallow networks are not sufficient to extract abundant features [41], [42], while deeper networks usually perform better. Wei et al. [19] presented a deep residual neural network for pansharpening. Yuan et al. [20] designed a multiscale and multidepth CNN by using convolution kernels of different sizes and two branch networks of different depths, in which features of different sizes and depths are fully utilized to achieve high pansharpening quality. Scarpa et al. [21] proposed a target-adaptive pansharpening method to ensure good results even in the case of mismatched training sets. Based on the idea of super-resolution, Hu et al. [22] proposed a two-stage pansharpening method, including a super-resolution stage and an image fusion stage. In the first stage, a residual network is used to improve the resolution of MS image. Then a multilevel detail injection network is designed in the second stage for image fusion.

Most of the above methods treat the CNN as a black box, without specific knowledge of pansharpening and lacking a explicit physical interpretation. For this problem, He et al. [23] and Deng et al. [24] adopted the detail injection framework of traditional methods to design networks. They exploited the nonlinear characteristics of CNNs to extract details, which were then injected into the upsampled MS image. Both methods are clearly interpretable and achieve fast convergence. Zhang et al. [29] proposed an network structure with two layers, two branches and two directions, which injects multiscale spatial details of PAN images into MS images in a layered and bidirectional manner, thus producing high spatial resolution output. According to the traditional MRA method, a block is designed to extract structure information from PAN image. And multiscale convolution kernel modules are also used to deepen and broaden the network.

Although these methods have greatly improved the pansharpened results, they mainly focus on the design of the depth of the network. The features extracted by the network are not fully and reasonably utilized. Motivated by the great success of attention mechanisms in various deep learning-related fields, such as image classification image segmentation, some researchers have introduced existing attention mechanisms to pansharpening tasks. Li et al. [25] designed a multiscale residual network by introducing the channel attention mechanism. Luo et al. [26] proposed a channel similarity attention fusion network by improving the channel attention module, which effectively suppressed redundant features. Rui et al. [27] and Wang et al. [28] use both channel attention and spatial attention modules to fully exploit useful features and suppress less useful ones to further improve pansharpening quality. In general, compared with traditional methods, CNN-based methods can better preserve spatial and spectral information.

C. GAN-Based Methods

Recently, GANs [30] have been successfully applied to various vision tasks due to their powerful image generation capabilities. Therefore, some works [31], [32], [33], [34], [35], [36], [37], [38] have also appeared to solve the pansharpening tasks by GANs. GANs have their own unique network structure. In generally, a GAN consists of a generator network and a discriminator network. Through continuous adversarial learning between the two, the generator eventually produces more realistic image samples that cannot be distinguished by the discriminator. Liu et al. [31] introduced GANs into the pansharpening field for the first time, generating HRMS images through a two-stream fusion generator network. Then they [33] proposed three different generator networks which are the extension of the previous [31]. Based on conditional GANs, Shao et al. [32] employed an encoder–decoder network with residual structure as the generator to avoid the loss of details caused by network deepening. Zhao et al. [36] applied fast guided filter and spatial attention mechanism to GANs to better preserve spatial information. In addition to designing the generator, some works improved GANs by using dual discriminators. Ma et al. [37] proposed a pansharpening method based on dual-discriminator network, which employs two discriminators to force the spectral and spatial information of the pansharpened image to be consistent with the LRMS and PAN images, respectively. Spectral information is discriminated from upsampled LRMS and pansharpened MS images, and spatial information is discriminated from average pooled pansharpened MS and PAN images. Gastineau et al. [38] also designed a GAN structure with dual discriminators, by using the intensity component and near-infrared band of pansharpened image for spatial information discrimination, and using Cr and Cb components for spectral information discrimination.

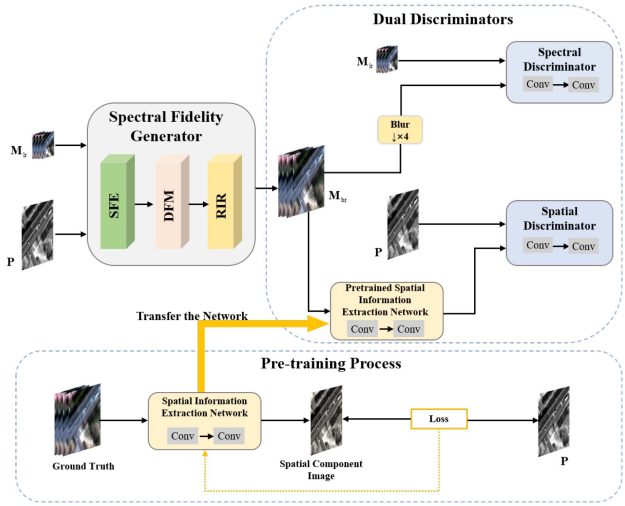

Fig. 2. Overall framework of the algorithm.

III. Proposed Method

A. Overview of the Framework

The overall framework of the proposed method is illustrated in Fig. 2, which consists of a spectral fidelity generator and dual discriminators. The spectral fidelity generator can generate the pansharpened image from the input PAN and LRMS images. Its network can be mainly divided into three parts: 1) the shallow feature extraction part (SFE), 2) the deep feature mapping part (DFM), and 3) the residual image reconstruction part (RIR). In the DFM, we design a cross-region similarity attention module to establish the relationship between similar features. Then the

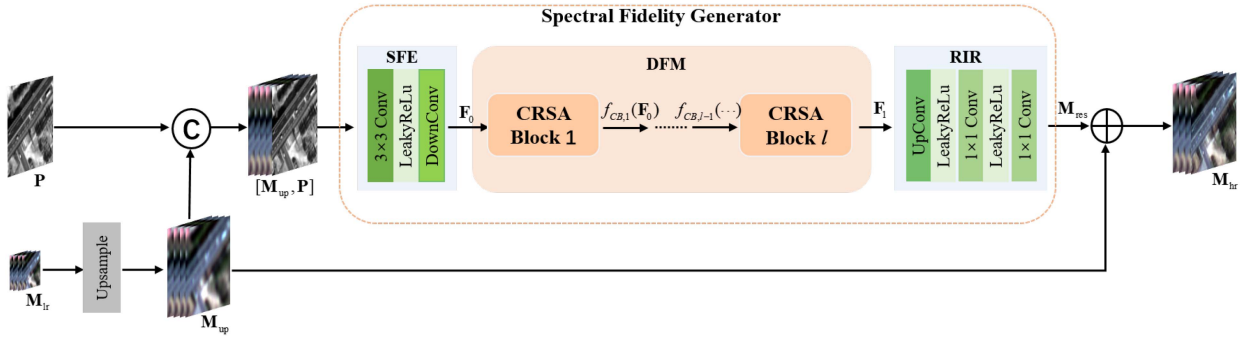

Fig. 3. Overall architecture of the spectral fidelity generator. **** and ⊕ denote concatenation operation and elementwise sum operation, respectively.

spatial component image of the pansharpened image is obtained through the pretrained spatial information extraction network. We pretrain the spatial information extraction network by using ground truth and PAN images. The ground truth is used as the input of the network, and the output image is compared with the PAN image to obtain the loss function to train the network. Finally, the network is capable of extracting spatial information of MS images. The pretrained network is then fixed in our dual-discriminator architecture to extract the spatial information of the pansharpened HRMS images. In other words, we utilize the invariant relationship between MS and PAN images, and transfer the relationship between ground truth and PAN image to the relationship between pansharpened HRMS image and its spatial component image. And spectral component image is obtained through blurring and downsampling. Finally, the spatial discriminator discriminates the spatial component image and PAN image, and the spectral discriminator discriminates the spectral component image and LRMS image. Through the continuous adversarial learning between the spectral fidelity generator and the dual discriminators, the generated image can preserve the spatial information of PAN image and spectral information of LRMS image. We are finally able to obtain the pansharpened image by fusing PAN image and LRMS image with the trained spectral fidelity generator.

B. Spectral Fidelity Generator

1) Network Architecture: The network architecture of the spectral fidelity generator is shown in Fig. 3. Let $\mathbf{M}_{\text{lr}} \in \mathbb{R}^{w\times h\times C}$ denote an LRMS image with $w \times h$ pixels and C spectral bands. Let $\mathbf{P} \in \mathbb{R}^{rw\times rh\times 1}$ denote a PAN image, where r is the spatial resolution ratio of $\mathbf{P}$ and $\mathbf{M}_{\text{lr}}$. First, $\mathbf{M}_{\text{lr}}$ is upsampled to the same size as the PAN image to obtain the upsampled MS image $\mathbf{M}_{\text{up}} \in \mathbb{R}^{rw\times rh\times C}$. Then $\mathbf{M}_{\text{up}}$ is concatenated with $\mathbf{P}$ along the spectral dimension to form the input of the network. The network starts with the SFE, which consists of two 3×3 convolutional layers with stride 1 and 2, respectively. The first convolutional layer is followed by a LeakyReLu activation function. The shallow feature $\mathbf{F}_0$ extracted by the SFE is denoted as

$$\mathbf{F}_0 = f_{\text{SFE}}([\mathbf{M}_{\text{up}}, \mathbf{P}]) \tag{1}$$

where $[\mathbf{M}_{\text{up}}, \mathbf{P}]$ is the concatenation of $\mathbf{M}_{\text{up}}$ and $\mathbf{P}$, and $f_{\text{SFE}}(\cdot)$ represents the convolution operation of the SFE. In order to obtain more expressive features during the fusion process, $\mathbf{F}_0$ is then fed into the DFM. DFM consists of l blocks, each of which is formed by embedding cross-region similarity attention module into residual dense block, called CRSA block. The process can be expressed as

$$\mathbf{F}_1 = f_{\text{DFM}}(\mathbf{F}_0) = f_{CB,l}(f_{CB,l-1}(\cdots f_{CB,1}(\mathbf{F}_0)\cdots)) \tag{2}$$

where $f_{\text{DFM}}(\cdot)$ indicates the function of the DFM, $f_{CB,l}(\cdot)$ is the function of the l_{th} CRSA block, and $\mathbf{F}_1$ is the final feature obtained by the DFM. The output of the DFM then enters the RIR to obtain the residual MS image. The RIR consists of one transposed convolutional layer and two 1×1 convolutional layers. This process can be expressed as

$$\mathbf{M}_{\text{res}} = f_{\text{RIR}}(\mathbf{F}_1) \tag{3}$$

where $f_{\text{RIR}}(\cdot)$ denotes the function of the RIR and $\mathbf{M}_{\text{res}}$ is the residual MS image after reconstruction by the generator network. In order to maintain the spectral information of the original MS image, we connect $\mathbf{M}_{\text{up}}$ to the output of the network for addition [18]. Finally, the pansharpened HRMS image $\mathbf{M}_{\text{hr}}$ is obtained, which is expressed as

$$\mathbf{M}_{\text{hr}} = \mathbf{M}_{\text{res}} + \mathbf{M}_{\text{up}}. \tag{4}$$

2) Cross-Region Similarity Attention Mechanism: Nonlocal operation can capture long-range dependencies by using matrix multiplication which computes the weighted sum of the features at all positions as the response of one feature [43]. Since matrix multiplication between feature maps can compute correlations among different positions, competitive performance has been achieved in areas, such as video classification and object detection. Based on the fact that a large number of pixels of the same substance in MS image have strong similarity in spectral and spatial features, matrix multiplication can be used to establish the relationship between them. Therefore, similar to [44], to obtain a mask through the attention mechanism and other operations, we introduce matrix multiplication to construct the cross-region similarity attention module, so as to solve the spectral distortion problem caused by the interaction of different substances at the junction.

The structure of cross-region similarity attention module is depicted in Fig. 4. Let $\mathbf{X} = [\mathbf{x}_{1,1}, \mathbf{x}_{1,2}, \ldots, \mathbf{x}_{i,j}, \ldots, \mathbf{x}_{\text{H,W}}]$ denote the input tensor with spatial size of H × W, where $\mathbf{x}_{i,j} \in \mathbb{R}^C$ represents the feature vector along the channel dimension at

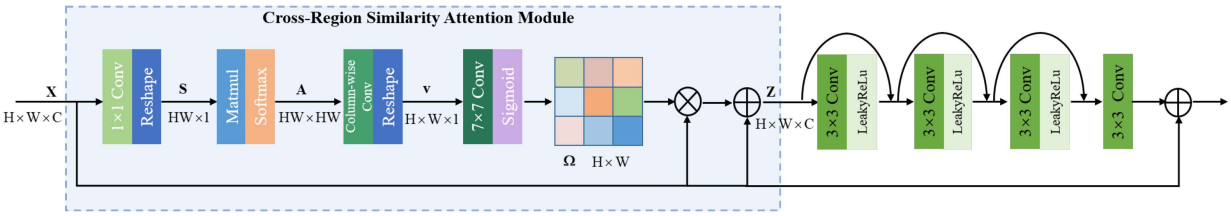

Fig. 4. Structure of the CRSA block, which integrates the cross-region similarity attention module into the residual dense block. ⊗ denotes the elementwise multiplication operation.

the spatial position (i, j). The channel dimension vector of each pixel in the MS image represents the spectral feature specific to the substance. So in order to preserve the spectral feature, we use 1×1 convolution to aggregate the channel dimension features to obtain a feature map in the spatial dimension, which can be expressed as

$$\mathbf{Y} = f_{\text{conv}}(\mathbf{X}) \tag{5}$$

where $\mathbf{Y} \in \mathbb{R}^{H \times W}$ is the resulting feature map. To calculate the interconnection between the features at each position, we first reshape $\mathbf{Y}$ to obtain a feature matrix $\mathbf{S}$ of size $\text{HW} \times 1$. Then we use the matrix multiplication to get the matrix $\mathbf{K}$, as follows:

$$\mathbf{K} = \mathbf{S} \cdot \mathbf{S}^T \tag{6}$$

where $\mathbf{S}^T$ is the transpose matrix of $\mathbf{S}$. After that, we use softmax to normalize each row of $\mathbf{K}$ to get a feature similarity matrix $\mathbf{A}$ of size $\text{HW} \times \text{HW}$. The value at position (i, j) of matrix $\mathbf{A}$ is formulated as

$$a_{i,j} = \frac{e^{k_{i,j}}}{\sum_{q=1}^{\text{HW}} e^{k_{i,q}}} \tag{7}$$

where $k_{i,j}$ denotes the value at position (i, j) of $\mathbf{K}$, and $a_{i,j}$ is the similarity value between the feature at position $(i, 1)$ in $\mathbf{S}$ and the feature at position $(j, 1)$ in $\mathbf{S}$. Each value in the ith row of $\mathbf{A}$ represents the similarity between the feature at position $(i, 1)$ in $\mathbf{S}$ and the feature at other position in $\mathbf{S}$. Based on the fact that the channel dimension of MS image represents the spectral information, the attention weights are desired to be added to the 2-D spatial dimension. To aggregate the feature similarity at the spatial position, we then perform a convolution operation on each column of $\mathbf{A}$ to get a global attention vector $\mathbf{u} = [\text{u}_1, \text{u}_2, \ldots, \text{u}_{\text{HW}}]$ of size $1 \times \text{HW}$, which is expressed as

$$\mathbf{u} = f_{\text{rconv}}(\mathbf{A}) \tag{8}$$

where $f_{\text{rconv}}(\cdot)$ represents the columnwise convolution. Then we reshape $\mathbf{u}$ to $\mathbf{v}$ of size $\text{H} \times \text{W} \times 1$, and each element $\text{v}_{i,j}$ in $\mathbf{v}$ represents the weighted similarity value between the position (i, j) and all other positions, which implies the interdependence between features at different positions. To learn the mask of cross-region similarity attention module from the similarity, we use a 7×7 convolutional layer and a sigmoid function $\sigma(\cdot)$ to $\mathbf{v}$, as follows:

$$\mathbf{\Omega} = \sigma(f_{\text{conv}}(\mathbf{v})) \tag{9}$$

where $\mathbf{\Omega} \in \mathbb{R}^{H \times W}$ is the mask in 2-D spatial dimension, which is then multiplied with the input feature. Finally, the output feature $\mathbf{Z}$ obtained by the input feature $\mathbf{X}$ after the cross-region similarity attention module is expressed as

$$\mathbf{Z} = \mathbf{X} + \mathbf{X} * \mathbf{\Omega} \tag{10}$$

where $*$ denotes the elementwise multiplication.

The proposed cross-region similarity attention mechanism can establish long-range dependencies and assign attention weights to features at different positions from a global perspective. To take full advantage of the cross-region similarity attention module, we further incorporate it into the residual dense block to form the CRSA Block, as shown in Fig. 4. The CRSA block contains four 3×3 convolutional layers, with LeakyReLu activation functions following all but the last layer. The output of each layer is used as the input of subsequent layers. According to the attention weights, the convolutional layers of the CRSA block will pay attention to the relationships between features at different positions, so that similar features can enhance each other. Moreover, the dense connection can make full use of all features and the residual structure can avoid gradient disappearance, which is beneficial to the network training.

C. Dual-Discriminator Architecture by Transferring Our Pretrained Spatial Information Extraction Network

As mentioned earlier, the spatial discriminator is responsible for distinguishing between the spatial component image of the pasharpened MS and the PAN image, and the spectral discriminator is responsible for distinguishing between the spectral component image of the pasharpened MS and the LRMS image.

1) *Spatial Discriminator:*

a) *Pretrained spatial information extraction network:* In order to get the spatial component image of the pansharpened MS image, we pretrain a spatial information extraction network with QB dataset. We use ground truth as the input image and PAN image as the reference image. The patch number and patch size of training and test images used are consistent with the settings of our training network finally. The role of the spatial information extraction network is to mine the potential relationship between the MS image domain and the PAN image domain. Through learning the relationship, the network achieves the extraction of the corresponding spatial component of MS image. Then it is transferred to our network architecture as an explicit functional module during network training and testing process. Compared with the average weighting of each band, our pretrained spatial information extraction network can fit the spatial component of the MS image more

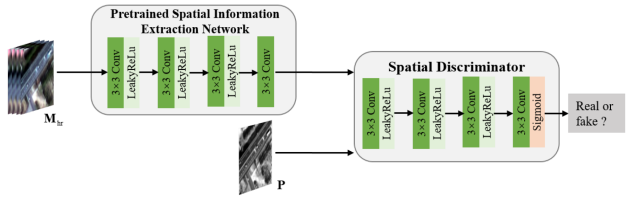

Fig. 5. Diagram of the discrimination of spatial information.

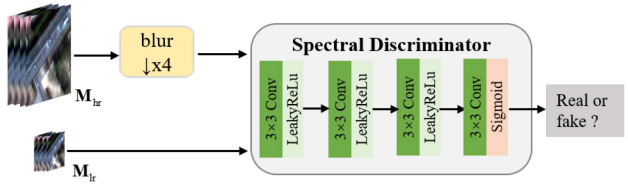

Fig. 6. Diagram of the discrimination of spectral information.

accurately by using the nonlinearity of convolution in an adaptive manner. As shown in Fig. 5, it consists of four 3×3 convolutional layers. Except for the last layer, the remaining layers are followed by the LeakyReLu function for nonlinear activation.

b) *Discrimination of spatial information:* The spatial component image obtained by the pretrained spatial information extraction network is then fed to the spatial discriminator. Similar to [33], we use a full convolutional network to construct our discriminators. The spatial discriminator consists of four 3×3 convolutional layers, and the number of convolution kernels in each layer is 32, 64, 128, and 1, respectively. Since the input of the spatial discriminator is the PAN image with a large size, we set the stride to 2 for the first two layers and 1 for the last two layers. All layers except the last one are followed by the LeakyReLu function. The last layer uses the sigmoid activation function to control the output between 0 and 1, thus outputting a discriminant result to distinguish true or false.

2) Spectral Discriminator: For the discrimination of spectral information, we first obtain the spectral component image with the same resolution as the LRMS image by blurring and downsampling the pansharpened image. This follows theWald's protocol [45], which processes the raw MS image in this way to obtain the LRMS image. This process also reflects the fact that the LRMS image has the spectral information of the raw MS image. The spectral component image is then fed into the spectral discriminator for discrimination. As shown in the Fig. 6, the structure of the spectral discriminator is similar to that of the spatial discriminator, which also contains four convolutional layers, but the stride of each layer is set to 1. Since the batch normalization layer is not suitable for low-level vision tasks [46], it is removed from our network.

D. Loss Functions

1) Loss Function of Spectral Fidelity Generator: Due to the dual-discriminator architecture, the loss function of the spectral fidelity generator includes a content-based loss and adversarial losses between itself and two discriminators, which is defined as

$$L_G(\theta_G) = \alpha L_c + \beta L_{\text{adv1}} + \gamma L_{\text{adv2}} \tag{11}$$

where $L_c, L_{\text{adv1}}, L_{\text{adv2}}$ represent the content-based loss, the adversarial loss with the spatial discriminator, and the adversarial loss with the spectral discriminator, respectively. α, β, and γ are the weight coefficients of each loss term, respectively.

For the content-based loss L_c, we adopt L_1 loss between the pansharpened image and ground truth, which is expressed as follows:

$$L_c = \frac{1}{N}\sum_{n=1}^{N} \|\mathbf{M}_{\text{ref}} - G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G)\|_1 \tag{12}$$

where N is the number of training samples in a minibatch, $\|\|_F$ is the Frobenius norm, $\mathbf{M}_{\text{ref}} \in \mathbb{R}^{rw \times rh \times C}$ indicates the ground truth, $G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G) = \mathbf{M}_{\text{hr}}$ is the pansharpened image and θ_G is the parameter of generator.

The adversarial loss L_{adv1} is represented as

$$L_{\text{adv1}} = \frac{1}{N}\sum_{n=1}^{N} \log(1 - D_1(S(G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_S); \theta_{D_1})) \tag{13}$$

where θ_{D_1} is the parameter of the spatial discriminator D_1, $S(\cdot)$ represents the function of spatial information extraction network and its parameter is θ_S, and $S(G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_S)$ is the spatial component of the pansharpened image.

The adversarial loss L_{adv2} is represented as

$$L_{\text{adv2}} = \frac{1}{N}\sum_{n=1}^{N} \log(1 - D_2(B \cdot R \cdot G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_{D_2})) \tag{14}$$

where θ_{D_2} is the parameter of the spectral discriminator D_2, B denotes blurring by Gaussian filter, R denotes downsampling, and $B \cdot R \cdot G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G)$ is the spectral component of the pansharpened image.

2) Loss Functions of Dual Discriminators: Until the spatial discriminator cannot distinguish the input spatial component image and the PAN image, and the spectral discriminator cannot distinguish the input spectral component image and the LRMS image, it can be considered that the discriminators achieve the training purpose. The corresponding loss function of the spatial discriminator D_1 is as follows:

$$\begin{aligned} L_{D_1}(\theta_{D_1}) = \frac{1}{N}\sum_{n=1}^{N}[&-\log(D_1(\mathbf{P}; \theta_{D_1})) \\ &- \log(1 - D_1(S(G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_S); \theta_{D_1}))] \end{aligned} \tag{15}$$

where $D_1(\mathbf{P}; \theta_{D_1})$ and $D_1(S(G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_S); \theta_{D_1})$ denote the classification probabilities of the PAN image and the spatial component image, respectively.

The loss function of the spectral discriminator D_2 is as follows:

$$\begin{aligned} L_{D_2}(\theta_{D_2}) = \frac{1}{N}\sum_{n=1}^{N}[&-\log(D_2(\mathbf{M}_{\text{lr}}; \theta_{D_2})) \\ &- \log(1 - D_2(B \cdot R \cdot G(\mathbf{M}_{\text{lr}}, \mathbf{P}; \theta_G); \theta_{D_2}))] \end{aligned} \tag{16}$$

where $D_2(\mathbf{M}_{\text{lr}};\theta_{D_2})$ and $D_2(B\cdot R\cdot G(\mathbf{M}_{\text{lr}},\mathbf{P};\theta_G);\theta_{D_2})$ are the classification probabilities of the LRMS image and the spectral component image, respectively.

IV. Experimental Results and Analyses

A. Datasets

There are four datasets used in our experiments, from QuikBird (QB), GaoFen-2 (GF-2), WorldView-2 (WV-2), and WorldView-3 (WV-3) satellite sensors. The spatial resolutions of the raw PAN images in the three datasets are 0.6, 0.8, 0.5, and 0.31 m, and the spatial resolutions of the corresponding MS images are 2.4, 3.2, 2.0, and 1.24 m, respectively. The ratio of spatial resolution is 4. The MS images contain four bands of red, green, blue, and near-infrared. Each dataset contains nine pairs of large-scale PAN and MS images, of which eight pairs for training and one pair for testing.

Since there are no available HRMS images as ground truth, we need to process the raw images according to the Wald's protocol [45]. We blur the raw PAN and MS images through a Gaussian filter, and downsample them by a factor of 4 to obtain the input low-resolution PAN and MS images required for the experiments. The raw MS images are then used as ground truth for comparison with the pansharpened images. In order to obtain sufficient training data, we crop the processed PAN and MS images into image patch pairs of size 128×128 and 32×32 by partially overlapping cropping, respectively. For the QB, GF-2, and WV-2 sensors, we obtained 19297, 24610, and 46208 training image patch pairs, respectively.

B. Evaluation Criteria

The performance evaluation is conducted both at reduced and full resolutions. In the reduced resolution experiments, we quantitatively evaluate the experimental performance with four widely used metrics: spectral angle map (SAM) [47], spatial correlation coefficient (SCC) [48], relative dimensionless global error in synthesis (ERGAS) [49], and universal image quality index Q for four-band images (Q4) [50]. These four metrics evaluate the similarity between the pansharpened image and ground truth. Specifically, SAM measures the spectral distortion by calculating the angle between the corresponding spectral vectors of the pansharpened image and the ground truth. A smaller SAM value indicates less spectral distortion. The ideal value of SAM is 0. SCC is used to measure the correlation of spatial information between the pansharpened image and the ground truth, and its ideal value is 1. The closer the SCC is to 1, the smaller the spatial distortion of the pansharpened image. ERGAS is a global quality evaluation metric, which evaluates the pansharpening quality globally through the root mean square error between two images in each band. The ideal value is 0. Q4 is the indicator that specifically measures four-band images, and its ideal value is 1.

Furthermore, to evaluate the performance of each method in full resolution experiments, we employ three widely used nonreference metrics, namely, D_λ, D_s, and QNR [51]. D_λ is a measure of spectral distortion based on the Q index between the pansharpened image bands. D_s is a metric to measure the spatial distortion based on the Q index between each band of the pansharpened image and PAN image. The ideal values of D_λ and D_s are 0. QNR is the abbreviation of quality with no reference, which is a global metric to measure the quality of pansharpened images without ground truth. It is a combination of the above two metrics and its ideal value is 1.

TABLE I
Quantitative Results of Different Methods on the QB Dataset

Method	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
GS [7]	2.1793	0.9532	2.2464	0.7934
BDSD [52]	2.8100	0.9343	2.6400	0.8040
PRACS [53]	2.7243	0.9286	2.3566	0.8310
AWLP [54]	3.1716	0.9037	3.2756	0.7491
SFIM [55]	1.9906	0.9412	2.6949	0.8120
MTF-GLP-HPM [56]	2.6815	0.8915	5.1931	0.7164
MTF-GLP-CBD [57]	2.8256	0.9228	2.7747	0.7917
PanNet [18]	1.3446	0.9914	1.1383	0.9912
FusionNet [24]	1.4238	0.9903	1.2069	0.9901
FU-PSGAN [33]	1.3972	0.9918	1.1105	0.9918
RED-cGAN [32]	1.3021	0.9922	1.1101	0.9916
Ours	**1.2844**	**0.9927**	**1.0792**	**0.9920**

The best result in each metric is in bold font.

C. Experimental Setup

Our method is implemented in Python3.6 with the Tensorflow1.3 framework and our experimental environment is based on the Ubuntu 18.04 operating system with the GPU NVIDIA-RTX 2080Ti. The spatial information extraction network is pretrained using the Adam optimizer with a learning rate of 0.0001. In both the generator and discriminators, the Adam optimizer is employed, and the initial learning rate is set to 0.0002. The minibatch and epoch are set to 16 and 50, respectively. In addition, the parameters of generator loss terms are set as: $\alpha=1$, $\beta=1\times10^{-4}$, $\gamma=1\times10^{-4}$.

D. Reduced Resolution Experiments

To evaluate the performance of our method, we compared it with some state-of-the-art methods, including three CS methods: GS [7], BDSD [52] and PRACS [53], and four MRA methods: AWLP [54], SFIM [55], MTF-GLP-HPM [56], and MTF-GLP-CBD [57], and three deep learning-based methods: PanNet [18], FusionNet [24], and FU-PSGAN [33]. Three different network structures are proposed in [33], among which FU-PSGAN has the best average performance on each dataset. Therefore, we choose FU-PSGAN as a comparison.

1) QB Dataset: First, we conduct a quantitative comparison of these methods on the QB dataset, as shown in Table I. We also compare the method named RED-cGAN [32] on this dataset. As can be seen from the table, our method can achieve the best results in all metrics among the compared methods, which can illustrate that our method has better fusion performance. This is because our spectral fidelity generator can effectively reduce the spectral distortion and our spatial information extraction network can help to recover more accurate spatial information.

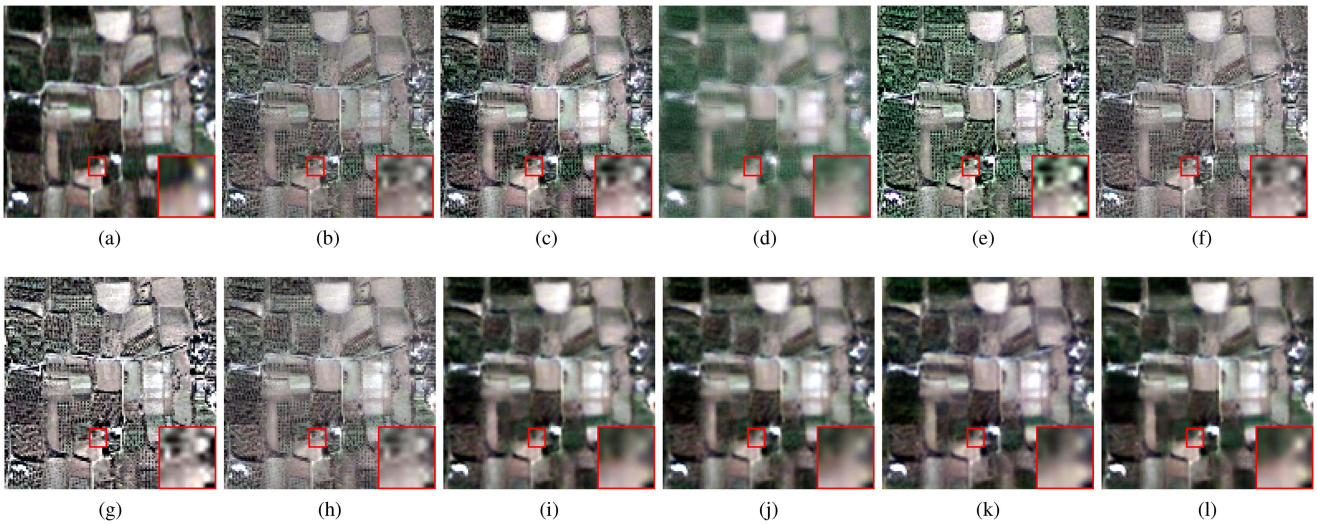

Fig. 7. Visual results obtained by different methods on the QB dataset. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE QB DATASET (SIZE OF THE PAN IMAGE: 256X256)

Method	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
PanNet [18]	1.345	0.9914	1.1375	0.9913
FusionNet [24]	1.4262	0.9904	1.2079	0.9902
FU-PSGAN [33]	1.3958	0.9918	1.1088	0.9919
Ours	**1.2777**	**0.9929**	**1.0611**	**0.9923**

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE GF-2 DATASET

Method	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
GS [7]	1.8379	0.9382	2.6864	0.7297
BDSD [52]	2.1426	0.9106	3.2766	0.7229
PRACS [53]	2.6997	0.9090	3.2535	0.7439
AWLP [54]	1.4114	0.9190	2.7054	0.7377
SFIM [55]	1.5516	0.9403	2.3681	0.8016
MTF-GLP-HPM [56]	2.0214	0.9078	3.3025	0.7218
MTF-GLP-CBD [57]	1.9854	0.9174	3.0269	0.7477
PanNet [18]	0.8771	0.9890	1.1311	0.9967
FusionNet [24]	0.8973	0.9886	1.2163	0.9961
FU-PSGAN [33]	0.8116	0.9917	0.9353	0.9977
Ours	**0.7755**	**0.9921**	**0.9050**	**0.9979**

Further from the results of the traditional methods and the deep learning-based methods, we can see that the latter performs far better than the former, which indicates the great advantage of deep learning. We also give experimental results on larger sized images, with PAN image size 256×256. The performance compared to other deep learning-based methods is shown in Table II. It can be seen that our method also works well on larger sized images.

For visual comparison, we further show the corresponding pansharpened images in RGB format in Fig. 7. Among them, Fig. 7(a) is the ground truth, and others are the visual results obtained by different pansharpening methods. It can be clearly observed that the results obtained by the AWLP and PRACS methods have obvious spectral distortion compared to the ground truth. The PRACS method blurs the whole image due to the lack of injected details, while some traditional methods seem to return images which look even sharper than the ground truth, with many more high-frequency details. Therefore, it seems to provide an enhanced version of the ground truth. However, it may be debatable whether this is desirable for pansharpening methods. Among the deep learning-based methods, FU-PSGAN also suffers from some spectral distortion, while others can preserve the color distribution well in the visual sense. As can be seen from the magnified region in the red box, our method is able to preserve more details, while others' details are missed, and thus, can only acquire poor quality. PanNet and FusionNet methods mainly focus on high-frequency information, ignoring the significance of low-frequency spatial information, which results in the loss of spatial details. However, our proposed spatial information extraction network can help to obtain a right amount of spatial information during the pansharpening process. Fig. 8 shows the absolute error maps (AEMs) obtained by taking a difference between the ground truth and the pansharpened results. The closer the color of the image is to black, the better the result. We can see that our result is closer to black in color. This is because our method can better preserve the spectral and spatial information through the spectral fidelity generator and dual discriminators, so the overall pansharpening quality is better than others.

2) GF-2 Dataset: The quantitative results of the different methods on the GF-2 dataset are shown in the Table III. From the results, we can see that our method obtains maximum values on both SCC and Q4 metrics compared to other methods. The SAM and ERGAS metrics are greatly reduced, which indicates that our cross-region similarity attention mechanism can take advantage of the complementary relationship between similar features to

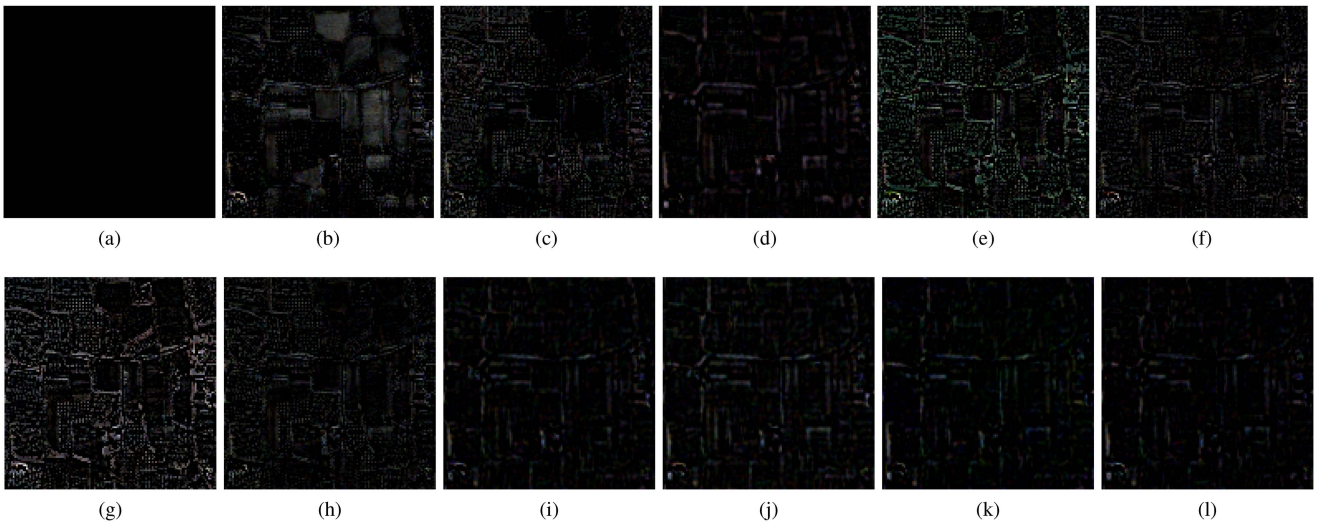

Fig. 8. AEMs of Fig. 7. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

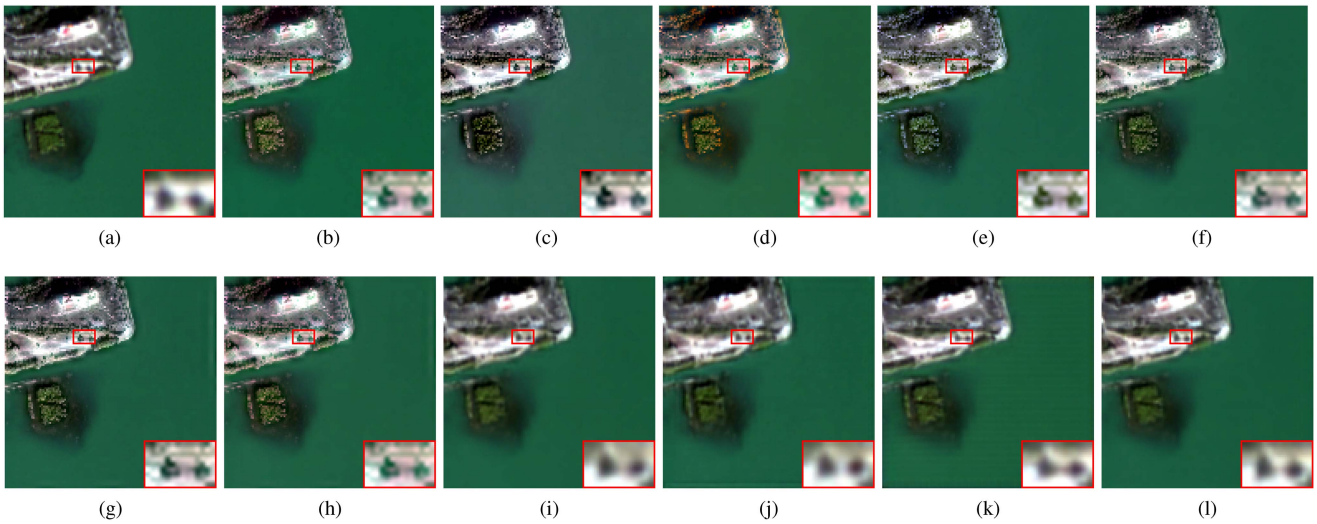

Fig. 9. Visual results obtained by different methods on the GF-2 dataset. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

effectively avoid spectral distortion and further improve the overall quality of the image.

In terms of visual comparison, the pansharpened results in Fig. 9 show that the CS and MRA methods have severe spatial distortion not only in most regions on the land, but also on the green plants on the lake, with a lot of noise that seriously affects the pansharpening quality. In addition, the result obtained by the BDSD method has obvious color changes on the lake surface. From the magnified region in the red box, it can be seen that the results of all methods except the deep learning-based methods have a large color difference with the ground truth, which illustrates the disadvantage of these methods in terms of spectral preservation. The results obtained by the GS, PRACS, and MTF-GLP-CBD methods are significantly brighter in color. Further observation of the spatial details from the local magnified region, it can be concluded that our method can better preserve the edges of the object and is slightly better than FU-PSGAN in reducing artifacts. This is because our method can obtain more accurate spatial information of pansharpened image through the spatial information extraction network, which makes it better in detail preservation than other methods. Also, it can be seen from the AEMs in Fig. 10 that different methods show different distortions on the land. Our method yields the smallest image error with the least distortion, which further demonstrate the effectiveness of our network.

3) WV-2 Dataset: As shown in Table IV, we show the quantitative results of each method on the WV-2 dataset. Similar conclusion can be drawn that the best performance is still reached by our method. Our method has a large improvement about the SAM metric, which can demonstrate the role of our proposed attention in spectral fidelity.

The visual comparison in Fig. 11 further corroborates the quantitative results. The AWLP, SFIM, MTF-GLP-HPM, and MTF-GLP-CBD methods intuitively suffer from severe color

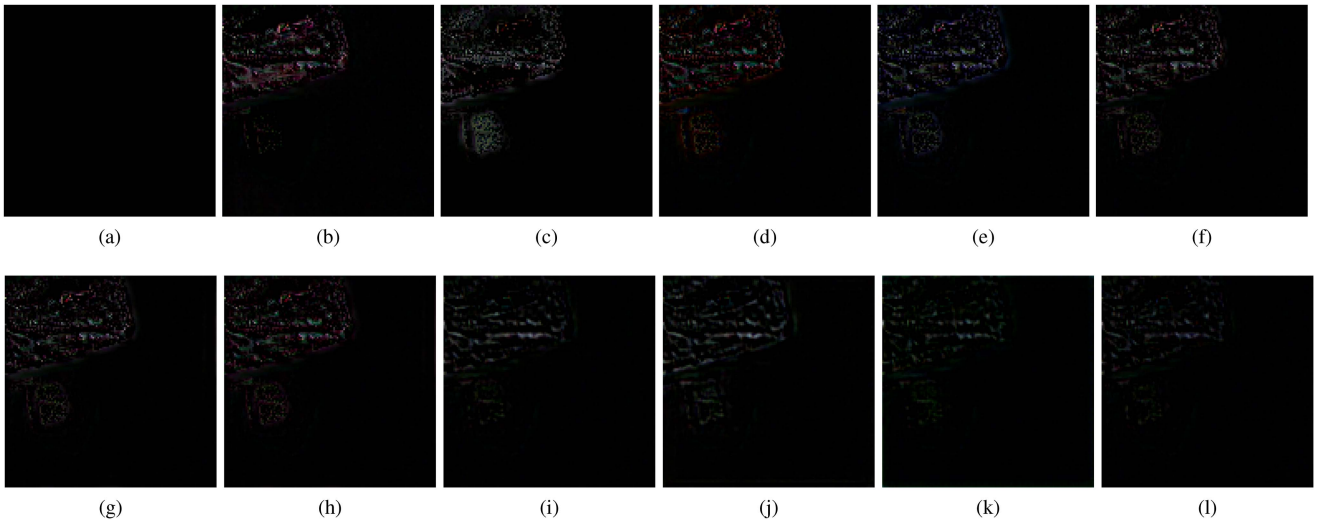

Fig. 10. AEMs of Fig. 9. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

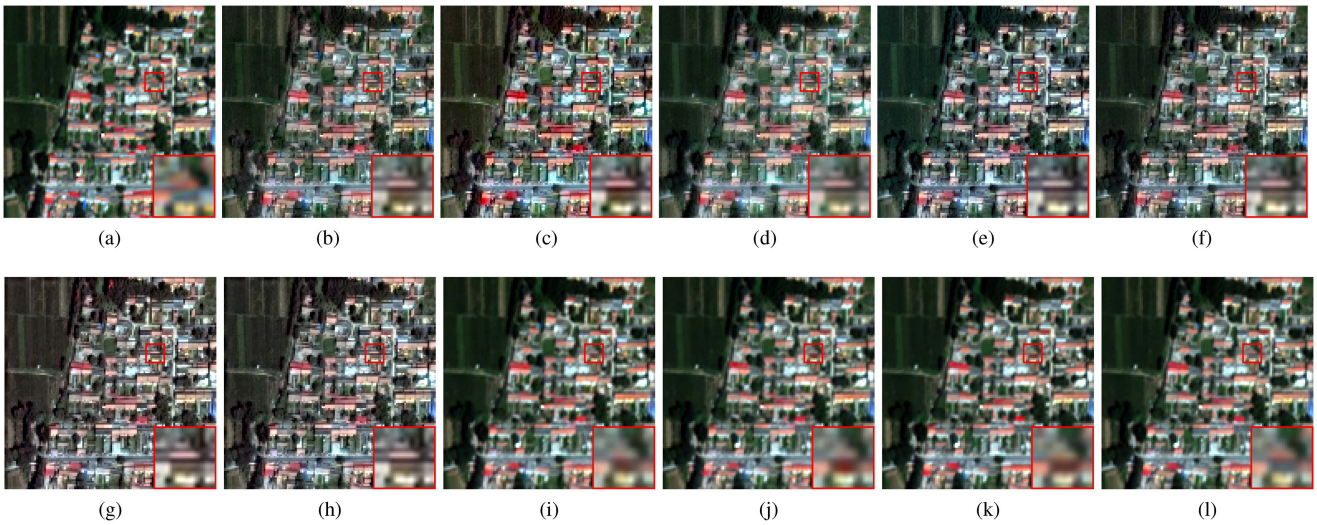

Fig. 11. Visual results obtained by different methods on the WV-2 dataset. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE WV-2 DATASET

Method	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
GS [7]	2.0758	0.9529	1.9705	0.6988
BDSD [52]	2.2811	0.9553	2.2134	0.7069
PRACS [53]	1.9434	0.9369	1.8925	0.7429
AWLP [54]	2.5706	0.9188	2.4968	0.6753
SFIM [55]	1.9286	0.9537	4.6455	0.7155
MTF-GLP-HPM [56]	2.2491	0.8898	8.8934	0.6528
MTF-GLP-CBD [57]	2.2372	0.9467	2.2590	0.6976
PanNet [18]	0.8881	0.9972	1.6780	0.9970
FusionNet [24]	0.8956	0.9973	1.6693	0.9970
FU-PSGAN [33]	0.8825	0.9976	1.5857	0.9973
Ours	**0.8390**	**0.9977**	**1.5593**	**0.9974**

distortion. The methods in the first row increase the saturation of the image and cannot maintain the spectral information of the image. Except for the PRACS method, the texture of the images obtained by the other methods in the first row is too sharp due to the excessive injection of detail information. Due to the independence from ground truth, poor results are obtained by these methods. The visual results of several deep learning-based methods are very close to the ground truth, and it is not easy to distinguish the specific visual differences among them. However, from the houses in the magnified region, it can be noticed that the results of PanNet, FusionNet, and FU-PSGAN have some color changes compared to the ground truth. Our method not only maintains the color distribution as much as possible, but also can effectively distinguish the boundaries of

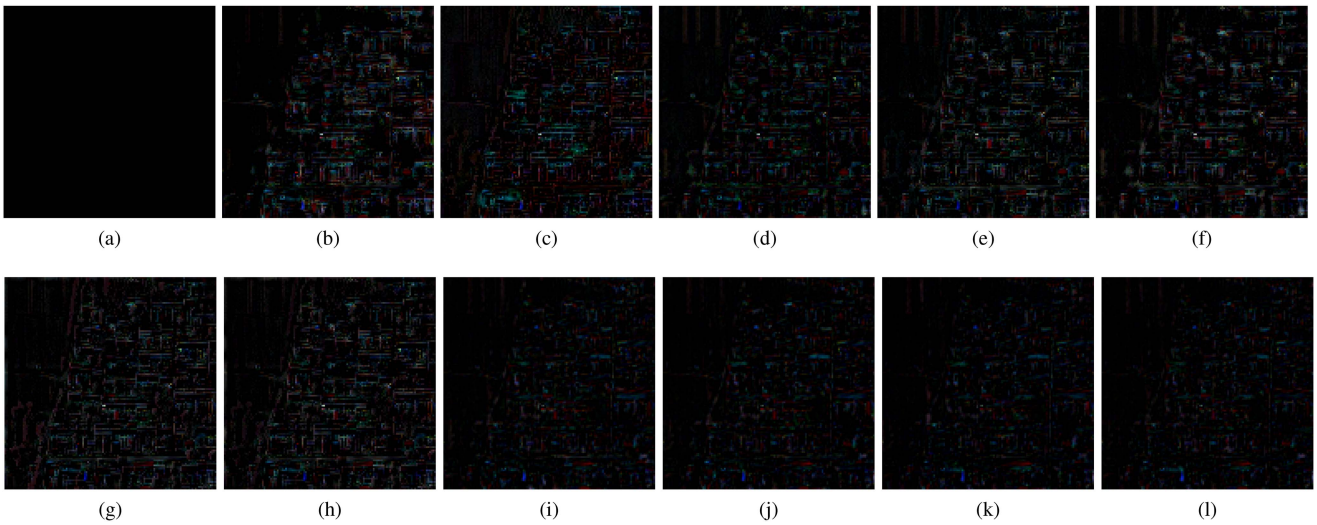

Fig. 12. AEMs of Fig. 11. (a) Ground Truth. (b) GS. (c) BDSD. (d) PRACS. (e) AWLP. (f) SFIM. (g) MTF-GLP-HPM. (h) MTF-GLP-CBD. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

TABLE V
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE WV-3 DATASET

Method	SAM(↓)	SCC(↑)	ERGAS(↓)	Q8(↑)
CDIF [15]	4.8376	0.9404	4.4572	0.8274
FusionNet [24]	3.3786	0.9797	2.4818	0.9026
FU-PSGAN [33]	3.4038	**0.9826**	2.5067	0.9017
Ours	**3.0685**	**0.9824**	**2.4687**	**0.9087**

TABLE VI
FULL RESOLUTION EXPERIMENTAL RESULTS OF DIFFERENT METHODS ON THE QB DATASET

Method	$D_\lambda(\downarrow)$	$D_s(\downarrow)$	QNR(↑)
GS [7]	0.0673	0.2796	0.6728
BDSD [52]	0.0473	0.2512	0.7138
PRACS [53]	0.1001	0.3394	0.5959
AWLP [54]	0.0941	0.2606	0.6721
SFIM [55]	0.0701	0.1654	0.7773
MTF-GLP-HPM [56]	0.1089	0.2928	0.6329
MTF-GLP-CBD [57]	0.0828	0.2397	0.6985
PanNet [18]	0.0142	**0.0307**	**0.9556**
FusionNet [24]	0.0184	0.0445	0.9379
FU-PSGAN [33]	0.0079	0.0485	0.944
Ours	**0.0059**	0.0477	0.9466

different colors without mixing. This fully demonstrates that our cross-region similarity attention mechanism can pay attention to the dependencies between similar substances, allowing them to promote each other, while attenuating the spectral distortions caused by neighboring different substances. Furthermore, consistent conclusions can also be drawn from the AEMs in Fig. 12 that our method brings about minimal spatial and spectral distortions.

4) WV-3 Dataset: The number of training and test images of WV-3 dataset is 9714 and 20, respectively. Since the competitive performance of FusionNet and FU-PSGAN, we choose these two methods as our compared ones in WV-3 dataset with eight spectral bands. A new method named by CDIF [15] was also selected to conduct experiments over WV-3 dataset. Performance indicators, such as SAM, SCC, ERGAS, and Q8 are employed to validate the efficiency of our proposed method.

From Table V, it can be seen that compared with the other three methods, our method can obtain competitive results. Only SCC which reflects the spatial quality, is slightly lower than that of FU-PSGAN method. For other three indicators, our method can achieve some improvement. The results show that our method can also be applied to datasets with different spectral and spatial resolutions. We can also see that CDIF performs poorly compared with other methods. This is because this method is a traditional one, which has an inherent gap compared with other deep learning methods. Another reason is that the method focuses on the relationship between PAN and HRMS images in the gradient domain and may ignore other key information needed for pansharpening, such as spectral information. And from Fig. 13, we can see that our method works well on WV-3 dataset and shows less image residuals compared with the other methods.

E. Full Resolution Experiments

To further validate the cross-scale generalization ability of our method, we also conduct experiments on full-scale images. We use the original MS and PAN images directly as input to the network optimized in the reduced resolution experiment, and evaluate the pansharpened images based on three nonreference metrics without ground truth. The quantitative experimental results are shown in Table VI. It can be seen that PanNet method is superior to other methods in both D_s and QNR. This is because PanNet method performs training in high-frequency domain, which makes it more flexible to be extended to full-scale images. We can observe that our method can obtain close performance to PanNet in global quality metric QNR, which proves the robustness of our method. We also present the qualitative results

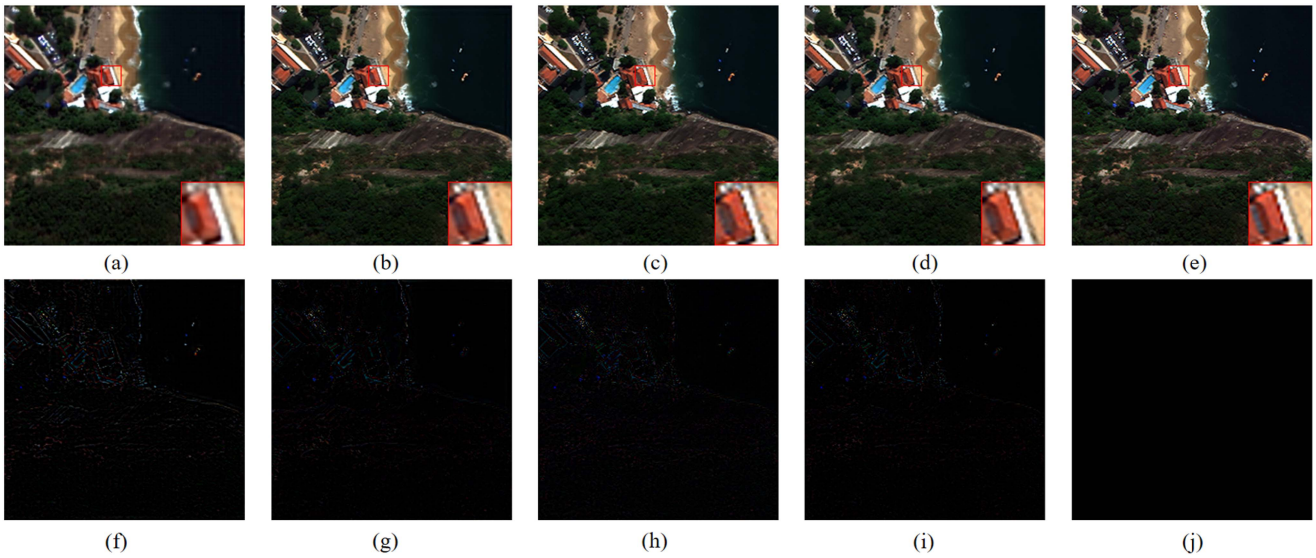

Fig. 13. Fusion results obtained by different methods on the WV-3 dataset. Visual results of (a) CDIF, (b) FusionNet, (c) FU-PSGAN, (d) Ours, and (e) Ground Truth, respectively. AEMs of (f) CDIF, (g) FusionNet, (h) FU-PSGAN, (i) Ours, and (j) Ground Truth, respectively.

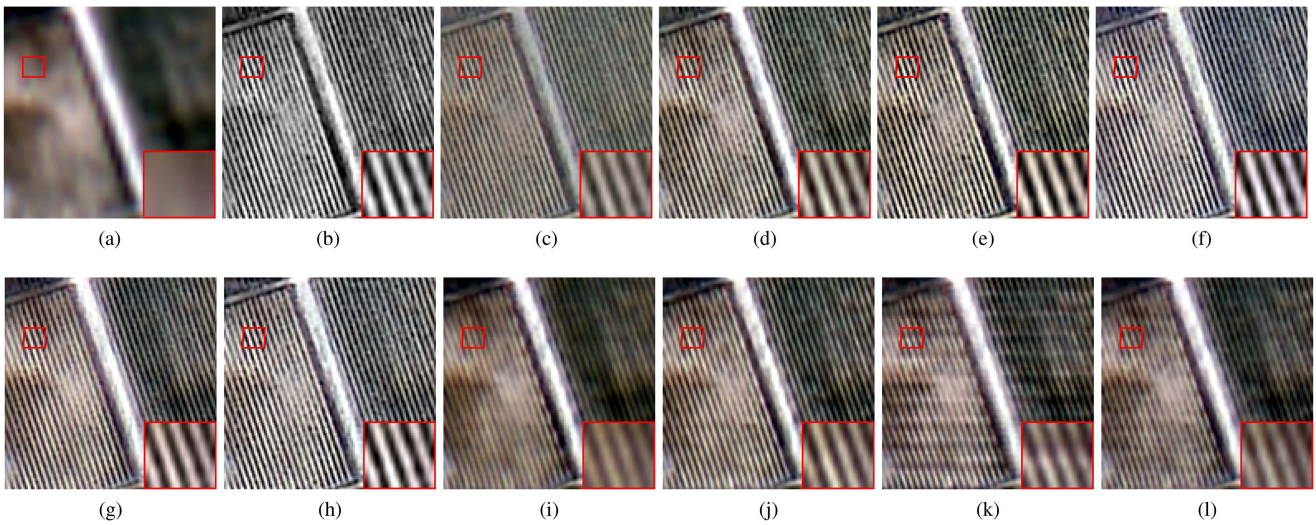

Fig. 14. Full-resolution results obtained by different methods on the QB dataset. (a) LRMS. (b) PAN. (c) GS. (d) BDSD. (e) PRACS. (f) AWLP. (g) SFIM. (h) MTF-GLP-HPM. (i) PanNet. (j) FusionNet. (k) FU-PSGAN. (l) Ours.

of each method in Fig. 14. It can be found that deep learning-based methods show better visual results and our method is most advantageous in spectral fidelity. This experimental result verifies the satisfactory performance of our method in terms of quantitative metrics and visual appearance.

F. Ablation Studies

1) Effect of the Number of CRSA Blocks: We first compare the performance of the network with different numbers of CRSA Blocks. As shown in Table VII, we set the number of blocks from 1 to 4, and obtain the following quantitative results. As can be seen from the table, the network performs better when the number of blocks increases from 1 to 2. However, as the number of blocks continues to increase, the performance gradually become worse. This is because the CRSA block contains the cross-region similarity attention module, which allows for not only the detailed features needed for low-level vision tasks but also contextually similar semantic features, compared to using

TABLE VII
COMPARISON WITH DIFFERENT NUMBERS OF CRSA BLOCKS

Number of blocks	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
1	1.3607	0.9914	1.1974	0.9906
2	**1.2844**	**0.9927**	**1.0792**	**0.9920**
3	1.3126	0.9921	1.0994	0.9918
4	1.3376	0.9912	1.1416	0.9912

TABLE VIII
COMPARISON OF DIFFERENT MODULE CONFIGURATIONS

Model	SAM(↓)	SCC(↑)	ERGAS(↓)	Q4(↑)
Baseline	1.3151	0.9922	1.0930	0.9920
Baseline+Spectral	1.3111	0.9921	1.0894	**0.9921**
Baseline+Spectral+Spatial	1.2979	0.9926	**1.0771**	0.9920
Baseline+Spectral+Spatial+Attention	**1.2844**	**0.9927**	1.0792	0.9920

only convolutional layers. When the number of blocks is too small, there are not enough similar dependencies to guide pansharpening for better performance. However, when the number of blocks is too large, the semantic features are too concentrated and important details may be lost. In addition, the depth of the network directly depends on the number of blocks. Increasing the number of blocks continuously may lead to overfitting of the model. Therefore, we adopt two CRSA blocks in the network eventually based on the experimental results.

2) Effect of Each Module: In order to verify the role of each module in the final network structure, the experimental performance of different module configurations is compared. Two discriminators and cross-region similarity attention module are removed from our network structure, and the generator-only structure is used as the baseline. Then, the spectral discriminator, spatial discriminator and attention module are added to the baseline in sequence. We conduct ablation experiments on the QB dataset and the results are shown in the Table VIII. As can be seen from the results, after adding the spectral discriminator to the baseline, both the SAM and ERGAS values are slightly reduced. Then after adding the spatial discriminator, it can be found that the SCC is improved. This is because the spatial extraction network can more accurately extract the spatial information to feed into the spatial discriminator, which is important for preserving the spatial information of the PAN images. At the same time, the values of other metrics are good, which indicates that the spatial and spectral discriminators combined with the generator to form adversarial learning can further improve the overall quality of pansharpened images. Finally, we add the cross-region similarity attention module. We can see that SAM is significantly reduced at this time. This is because the proposed attention mechanism can establish relationships between similar features, which is extremely helpful to avoid spectral distortion. Satisfactory results can be obtained in the final network structure, so the effectiveness of the modules used in combination can be illustrated.

V. CONCLUSION

In this article, a cross-domain association mining-based GAN is considered for pansharpening. To strengthen the inherent relationships between image pixels, our designed cross-region similarity attention extracts the dependencies between similar features by using matrix multiplication to effectively avoid spectral distortion. Then, we further constrain the spatial and spectral components of the pansharpened MS image through a spatial discriminator and a spectral discriminator. The pretrained spatial information extraction network is proposed to ensure that the spatial information of the input PAN image is preserved by exploiting the relationship between MS images and PAN images. Thereby, the pansharpening quality is improved. Experiments on three datasets show that our method can achieve excellent performance compared with several existing CS methods, MRA methods and deep learning-based methods, and can obtain the pansharpened image much closer to the ground truth.

Our method is the supervised pansharpening method, which requires manual training data and takes the original MS image as ground truth. As we all know, unsupervised pansharpening method is more suitable for real pansharpening scenes. Therefore, for real implementation, our future work will focus on unsupervised pansharpening method, which just uses original PAN and MS images as the input.

REFERENCES

[1] P. Shanmugapriya, S. Rathika, T. Ramesh, and P. Janaki, "Applications of remote sensing in agriculture—A review," *Int. J. Curr. Microbiol. Appl. Sci.*, vol. 8, no. 1, pp. 2270–2283, 2019.
[2] R. R. Girija and S. Mayappan, "Mapping of mineral resources and lithological units: A review of remote sensing techniques," *Int. J. Image Data Fusion*, vol. 10, no. 2, pp. 79–106, 2019.
[3] J. Wang, F. Li, and H. Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707013.
[4] C. Zhang, L. Wang, S. Cheng, and Y. Li, "SwinSUNet: Pure transformer network for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5224713.
[5] W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogrammetric Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, 1990.
[6] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.
[7] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875, 2000.
[8] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "Mtf-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogramm Eng. Remote Sens.*, vol. 72, pp. 591–596, 2006.
[9] L. Alparone, S. Baronti, B. Aiazzi, and A. Garzelli, "Spatial methods for multispectral pansharpening: Multiresolution analysis demystified," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2563–2576, May 2016.
[10] R. King and J. Wang, "A wavelet based algorithm for pan sharpening landsat 7 imagery," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2001, vol. 2, pp. 849–851.
[11] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, 2007.
[12] A. Garzelli, F. Nencini, L. Alparone, and S. Baronti, "Multiresolution fusion of multispectral and panchromatic images through the curvelet transform," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2005, vol. 4, pp. 2838–2841.
[13] M. Ghahremani and H. Ghassemian, "Remote-sensing image fusion based on curvelets and ICA," *Int. J. Remote Sens.*, vol. 36, no. 16, pp. 4131–4143, 2015.
[14] M. González-Audícana, J. L. Saleta, R. G. Catalán, and R. García, "Fusion of multispectral and panchromatic images using improved IHS and PCA mergers based on wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 42, no. 6, pp. 1291–1299, Jun. 2004.
[15] J. Xiao, T. Huang, L. Deng, Z. Wu, and G. Vivone, "A new context-aware details injection fidelity with adaptive coefficients estimation for variational pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2022.
[16] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, "Remote sensing image fusion with convolutional neural network," *Sens. Imag.*, vol. 17, no. 1, pp. 140–155, 2016.

[17] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.

[18] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "Pannet: A deep network architecture for pan-sharpening," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1753–1761.

[19] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[20] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[21] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[22] J. Hu, C. Du, and S. Fan, "Two-stage pansharpening based on multi-level detail injection network," *IEEE Access*, vol. 8, pp. 156442–156455, 2020.

[23] L. He et al., "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

[24] L. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6995–7010, Aug. 2021.

[25] X. Li et al., "A remote-sensing image pan-sharpening method based on multi-scale channel attention residual network," *IEEE Access*, vol. 8, pp. 27163–27177, 2020.

[26] S. Luo, S. Zhou, and Y. Qi, "CSAFNet: Channel similarity attention fusion network for multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2020, Art no. 5000205.

[27] X. Rui, Y. Cao, Y. Kang, W. Song, and R. Ba,, "Maskpan: Mask prior guided network for pansharpening," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 853–857.

[28] P. Wang and E. Sertel, "Channel–spatial attention-based pan-sharpening of very high-resolution satellite images," *Knowl.-Based Syst.*, vol. 229, 2021, Art. no. 107324.

[29] T. Zhang, L. Deng, T. Huang, J. Chanussot, and G. Vivone, "A triple-double convolutional neural network for panchromatic sharpening," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: 10.1109/TNNLS.2022.3155655.

[30] I. Goodfellow et al., "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 1–9.

[31] X. Liu, Y. Wang, and Q. Liu, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 873–877.

[32] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, and Y. Zhang, "Residual encoder–decoder conditional generative adversarial network for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1573–1577, Sep. 2020.

[33] Q. Liu, H. Zhou, Q. Xu, X. Liu, and Y. Wang, "PSGAN: A generative adversarial network for remote sensing image pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10227–10242, Dec. 2021.

[34] L. Zhang, W. Li, C. Zhang, and D. Lei, "A generative adversarial network with structural enhancement and spectral supplement for pan-sharpening," *Neural Comput. Appl.*, vol. 32, no. 24, pp. 18347–18359, 2020.

[35] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "A residual dense generative adversarial network for pansharpening with geometrical constraints," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 493–497.

[36] Z. Zhao et al., "FGF-GAN: A lightweight generative adversarial network for pansharpening via fast guided filter," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6.

[37] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "PAN-GAN: An unsupervised pan-sharpening method for remote sensing image fusion," *Inf. Fusion*, vol. 62, pp. 110–120, 2020.

[38] A. Gastineau, J.-F. Aujol, Y. Berthoumieu, and C. Germain, "Generative adversarial network for pansharpening with spectral and spatial discriminators," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[39] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[41] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Comput. Sci.*, 2014.

[43] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[44] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M. H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5728–5739.

[45] L. Wald, "Quality of high resolution synthesised images: Is there a simple criterion?," in *Proc. Fusion Earth data, Merging Point Meas., Raster Maps, Remotely Sensed Images*, 2000, pp. 99–103.

[46] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2017, pp. 136–144.

[47] R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.

[48] J. Zhou, D. L. Civco, and J. Silander, "A wavelet transform method to merge landsat TM and spot panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.

[49] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses Des MINES, 2002.

[50] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.

[51] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, 2008.

[52] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[53] 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.

[54] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

[55] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, 2000.

[56] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 930–934, May 2014.

[57] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, "Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.