


Cross Spectral and Spatial Scale Non-local Attention-Based Unsupervised Pansharpening Network

Shuangliang Li , Yugang Tian , *Member, IEEE*, Cheng Wang, Hongxian Wu, and Shaolan Zheng

Abstract—Pansharpening means fusing the low spatial resolution multispectral image (LRMSI) and the panchromatic (PAN) image to get the high resolution multispectral image (HRMSI). Due to the powerful feature learning ability of the deep-learning (DL), DL-based unsupervised fusion methods have been developed explosively. However, most of the fusion methods are difficult to fully explore and utilize the correct spatial and spectral correlation between the LRMSI, HRMSI, and PAN images. In addition, the CNN-dominated fusion framework is limited by its local feature learning without exploring the global feature dependency to further enhance the image feature. Therefore, to fully exploit the correct correlations between LRMSI, HRMSI, and PAN images and to explore the global feature dependency, we designed a cross-scale unsupervised fusion network (CSFNet). This network is composed of two cross spectral and spatial scale's nonlocal attention blocks to effectively fuse the LRMSI and PAN image features. And the fusion strategy is implemented by mapping the computed nonlocal similarity from the low resolution scale to the high resolution scale and outputs the reconstructed HRMSI feature. The experimental results on two datasets show that it achieves state-of-the-art performance compared to other fusion methods.

Index Terms—Cross scale, deep learning, nonlocal attention, pansharpening, unsupervised training.

I. INTRODUCTION

MULTISPECTRAL image (MSI) generally refers to the satellite image with the rich spectral information. It has been widely used in many applications, including land cover classification, change detection, and object recognition [1], [2]. However, due to the limitation of the imaging sensors, their low spatial resolution limits their application range and accuracy. Therefore, the coupled PAN image with the higher spatial resolution is always fused with LRMSI to get HRMSI with higher

spatial resolution, which is called the pansharpening [3], [4] [5], [6], [7].

Many approaches have been developed to fuse the LRMSI and PAN images, which can be summarized into four categories, including component substitute (CS)-based, multiresolution analysis (MRA)-based, variational optimization (VO), and deep learning (DL)-based. CS-based methods substitute the simulated intensity band from the LRMSI by the histogram-matched PAN image to improve the spatial quality of the LRMSI, mainly including intensity-hue-saturation (IHS) [8] and Gram-Schmidt adaptive (GSA) [9], [10]. MRA-based methods decompose the PAN image into different spatial scales and inject the corresponding high-frequency details into the LRMSI, including smoothing filter-based intensity modulation (SFIM) [11], wavelet transform (Wavelet) [12], modulation transfer function with generalized Laplacian pyramid (MTF_GLP) [23], and MTF_GLP with high-pass modulation (MTF_GLP_HPM) [23]. VO-based methods usually design the specific constrained target function and alternatively optimize the variables to achieve the best fusion performance. Among them, coupled nonnegative matrix factorization (CNMF) [23] is the most representative one. However, CS-based and MRA-based methods always introduce the spectral and spatial distortion into the fused results, while VO-based methods suffer from a high computational cost.

In recent years, DL-based methods have been widely explored due to their powerful feature learning ability [16], [17], [18]. Especially the convolutional neural networks (CNN), which can capture the complex local spatial and spectral features of the image, have been widely used in the image fusion field. For example, Masi et al. [16] first proposed a CNN-based pansharpening network-PNN, which includes three convolutional layers. Then, Yang et al. [17] developed a deep network structure-Pannet with a high-pass filtering method to fuse the LRMSI and PAN image. Kwan et al. [19] summarized 11 pansharpening algorithms to enhance the images of the left imager in the mastcam by its right imager. Li et al. [20] designed a detail injection network to inject the pan details into the LR image by rectifying the incorrect data distribution, which achieves great fusion performance.

Actually, the training stage of the above supervised methods is on the low resolution dataset, while the testing is on the high resolution scale. However, it is laborious and difficult to generate the low resolution dataset that could maintain the scale-invariant property of the image feature. As a result, this kind of training strategy may degrade the fused image quality

Manuscript received 15 February 2023; revised 31 March 2023 and 16 May 2023; accepted 17 May 2023. Date of publication 22 May 2023; date of current version 5 June 2023. This work was supported by the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources under Grant 2022NRM003. (*Corresponding author: Yugang Tian.*)

Shuangliang Li, Yugang Tian, and Cheng Wang are with the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources, Guangzhou 510620, China, and also with the School of Geography and Information Engineering, China University of Geosciences, Wuhan 430074, China (e-mail: cug_lsl@cug.edu.cn; ygangtian@cug.edu.cn; 20171003562@cug.edu.cn).

Hongxian Wu and Shaolan Zheng are with the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources, Guangzhou 510620, China (e-mail: 819289751@qq.com; shaolan51@163.com).

Digital Object Identifier 10.1109/JSTARS.2023.3278296

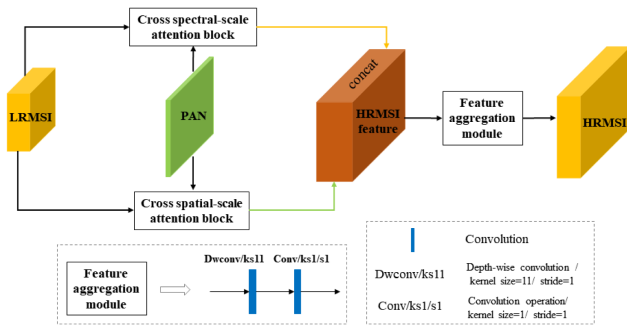


Fig. 1. Overall framework of the proposed fusion method. ‘concat’ means the bandwise concatenation operation on the image features generated from two cross-scale attention blocks.

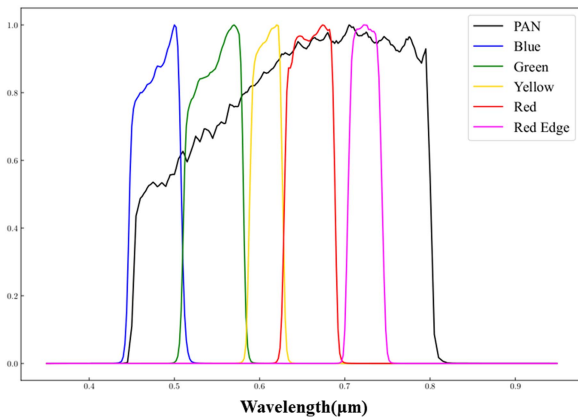


Fig. 2. Different bands’ spectral response functions (SRF) of the WorldView2 satellite imaging sensor.

on the original HR scale. Therefore, the unsupervised fusion network that performs the training process on the high resolution scale has been gradually developed in recent years. For example, Ma et al. [21] designed an unsupervised fusion network–PanGAN with spectral and spatial discriminators to get the fused HRMSI. Zhou et al. [22] proposed an unsupervised generative adversarial framework on the original high resolution scale. This framework extracts the modality-specific features from the LRMSI and PAN images with the designed generator and then fuse them to get the HRMSI. In addition, based on the self-attention mechanism, Qu et al. [23] propose an unsupervised pansharpening method in a deep-learning framework that achieves great performance.

However, these unsupervised fusion networks are trained by relying on the commonly used loss functions. For the remote sensing image, the space-invariant degradation kernels (e.g., “bicubic”) are always used to constrain the spectral fidelity of the fused HRMSI [21]. And the spatial constraint is to enforce the consistency between the PAN image and the bandwise average of the fused HRMSI. But these two constraints in the spectral and spatial domain may not be totally correct and lack the robustness for different image features. For example, as shown in Fig. 2, the spectral response function (SRF) of the MSI and PAN image bands do not fully follow the linear relationship,

which demonstrates the irrationality of the “bandwise average” consistency. Therefore, the fusion network that can correctly learn and explicitly use the spatial and spectral correlations is essential to achieve excellent fusion performance.

Furthermore, most of the fusion networks are constructed based on the widely used CNN structure. But the limited local feature learning of the CNN hinders the exploitation of the long-range dependence of the image features. Recently, the transformer structure [24] has been developed in computer vision fields, such as super-resolution [25], [26]. Its main idea is to use nonlocal self-attention to enhance the image feature. In the image fusion field, several nonlocal attention-based methods have also been proposed and achieved great performance. For example, Zhang et al. [27] proposed a multiscale spatial-spectral interaction transformer to complete the pansharpening task. Wele [28] adopted the transformer structure to inject the PAN image details into the hyperspectral image. However, these methods compute the nonlocal attention on the pixel level and may not sufficiently learn the local texture information of the image feature, which may degrade the resulting image quality.

To effectively solve the above problems, we make several contributions, which can be summarized as follows.

First, we design a cross spectral-scale nonlocal attention module. This module could map the spectral correlation computed between the LRMSI and PAN image features to the high resolution scale and reconstruct the HRMSI feature. Without using the “average” function to constrain the spatial fidelity of the fused HRMSI, we explicitly learn and exploit the spectral band correlation through the designed network structure.

Second, to sufficiently adapt to the spatial resolution difference between the LRMSI and PAN images, we propose a cross spatial-scale nonlocal attention module. This module first computes the patch-level nonlocal similarity between the encoded LRMSI and PAN images. Then we map this similarity to the original LRMSI feature to reconstruct the HRMSI feature.

Finally, compared with the other nonlocal attention-based fusion methods, the nonlocal similarity matrix in the designed cross spatial-scale block is computed on the patch level. The patch-level computation could greatly reserve the local texture and context information. And the nonlocal similarity matrix could capture the global attention information. This combination of local and nonlocal is much more effective and efficient in enhancing the image feature learning and improving the fusion performance.

The rest of this article is organized as follows. Section II introduces the related work with the research issues, including unsupervised pansharpening, patch recurrence property and nonlocal attention. Section III describes the designed network modules and loss functions. The results of the comparative experiments are given in Section IV. Section V shows the experimental result of the ablation study. Finally, Section VI concludes this article.

II. RELATED WORK

A. Unsupervised Pansharpening

Actually, traditional methods including CS, MRA, and VO-based all belong to unsupervised pansharpening methods.

However, due to the poor performance of these methods in spatial and spectral fidelity or huge computational cost, DL-based methods have been explored in recent years. The DL-based unsupervised method takes the training process on high resolution scale and could use appropriate loss functions to achieve satisfactory fusion performance.

For example, Zhou et al. [29] proposed a novel pansharpening framework that adopts the auto-encoder and perceptual loss to complete the fusion process. Qu et al. [23] proposed an unsupervised deep-learning framework based on the self-attention mechanism and achieved great fusion performance. In addition, Seo et al. [30] designed an unsupervised learning framework with registration learning for pansharpening and designed two novel loss functions to train the network. Diao et al. [31] designed a multiscale fusion network and adopted multiple GAN structures at different scales to improve the fusion performance. Xu et al. [32] proposed an iterative network based on spectral and texture loss constraints and generative adversarial network to fuse the LRMSI and PAN images. Although the great fusion performance they achieved, the lack of exploring and utilizing the correct spectral and spatial correlations is still the obstacle to achieving a better fusion performance.

B. Patch Recurrence Property and Nonlocal-Based Attention

Image patches tend to recur within and across scales of a same image, which is called the patch recurrence property [33], [34]. This property has been widely used in the image super-resolution field to enhance the HR patch using similar patches in images of different resolutions. In the pioneering study, Glasner et al. [33] integrated the methods of multiple images SR and example-based SR to exploit repeating patches within and across multiscale images. Furthermore, Freedman et al. [35] effectively extracted the patches from the localized regions of the input image, which could reduce the computational complexity.

This patch recurrence property within and across scales represents the long-range dependence of image patches, which is similar to the idea of the transformer network–nonlocal self-attention [24]. Some studies have used the nonlocal attention module to super-resolve the LR image. For example, Liu et al. [36] propose a nonlocal recurrent network to incorporate the nonlocal operations into a recurrent neural network for image restoration with fewer parameters. Dai et al. [37] proposed a channel attention module to adaptively rescale the channelwise features and a nonlocally enhanced residual group to capture long-distance spatial information. Different from the nonlocal self-attention in the super-resolution field, we need to consider the cross-attention between different resolutions' images in the fusion field. In addition, the nonlocal attention in some methods is computed on the pixel level, which may not be sufficient to learn the local texture feature of the image.

III. METHOD

A. Overall Fusion Framework

The proposed fusion framework is shown in Fig. 1. We design two cross spectral-scale and cross spatial-scale blocks to

sufficiently learn the spectral and spatial relations of two input images. These relations are then mapped to high resolution scale to reconstruct the HRMSI features. Then a “Feature aggregation module” is designed to integrate the HRMSI features from these two cross scale blocks and output the fused HRMSI.

B. Cross Spectral-Scale Nonlocal Attention Block

Actually, in some SRF (the mapping function from MSI to PAN image) estimation methods, they use the downsampled images to estimate the SRF and apply it on the high resolution scale [38], which achieves great performance. Therefore, the spectral band correlation on the LR scale is approximately the same as that on the HR scale. This could be expressed as

$$\psi(P_{h \times w}, M_{h \times w}^j) = \psi(\mathbb{P}_{H \times W}, \mathbb{M}_{H \times W}^j) \quad (1)$$

where $P_{h \times w}$ represents the low resolution PAN image, $M_{h \times w}^j$ represents the low resolution MSI image and $\mathbb{M}_{H \times W}^j$ represents the j th band of the high resolution MSI. Their spatial sizes are $h \times w$ and $H \times W$ on the LR and HR scales, respectively. $\psi(\cdot, \cdot)$ means the function to calculate the correlation as

$$\psi(P_{h \times w}, M_{h \times w}^j) = \theta(P_{h \times w}) \cdot \delta(M_{h \times w}^j) \quad (2)$$

where $\theta(\cdot)$ and $\delta(\cdot)$ are feature transformation functions. And \cdot means the dot-product operation to output the correlation coefficient between $P_{h \times w}$ and $M_{h \times w}^j$.

However, most methods only use the estimated spectral band correlation to constrain the spatial fidelity in the loss function part, which may not make full use of it. Therefore, we design a cross spectral-scale nonlocal attention block to explicitly map the computed bandwise nonlocal similarity from the LR scale to the HR scale and reconstruct the HRMSI feature, as shown in Fig. 3. In this block, we first downsample both input images by a single stride convolution from $H \times W$ and $h \times w$ to the same size $\rightarrow h' \times w'$ (we set h', w' equal to $h/(r^2), w/(r^2)$). This downsampling operation could get the discriminative features and reduce the complexity of the following similarity computation. Note that we maintain the number of original spectral bands in this block. This unchanged spectral dimension could greatly preserve the spectral bands' self-correlation and cross-correlation of the input images.

Then we compute the bandwise similarity matrix

$$\phi(P_{1 \times h' \times w'}, M_{C \times h' \times w'}) = (P_{1 \times h' \times w'}) (M_{C \times h' \times w'})^T \Rightarrow S_{1 \times C} \quad (3)$$

where h' and w' represent the spatial size of the downsampled image features, as shown in the middle part of Fig. 3. “1” and “C” are the number of bands of PAN image and MSI, respectively. “ $S_{1 \times C}$ ” represents the computed nonlocal similarity matrix. Note that $\psi(\cdot, \cdot)$ in (2) is different from the $\phi(\cdot, \cdot)$ in (3). The former is computed on the single-band image while the latter is computed on the all-band image.

To map the similarity matrix computed on the LR scale to the HR scale, as shown in the right part of Fig. 3, we perform the matrix multiplication operation between the similarity matrix and PAN' to obtain the reconstructed HRMSI feature. Note that we first align the feature space of the PAN image by the

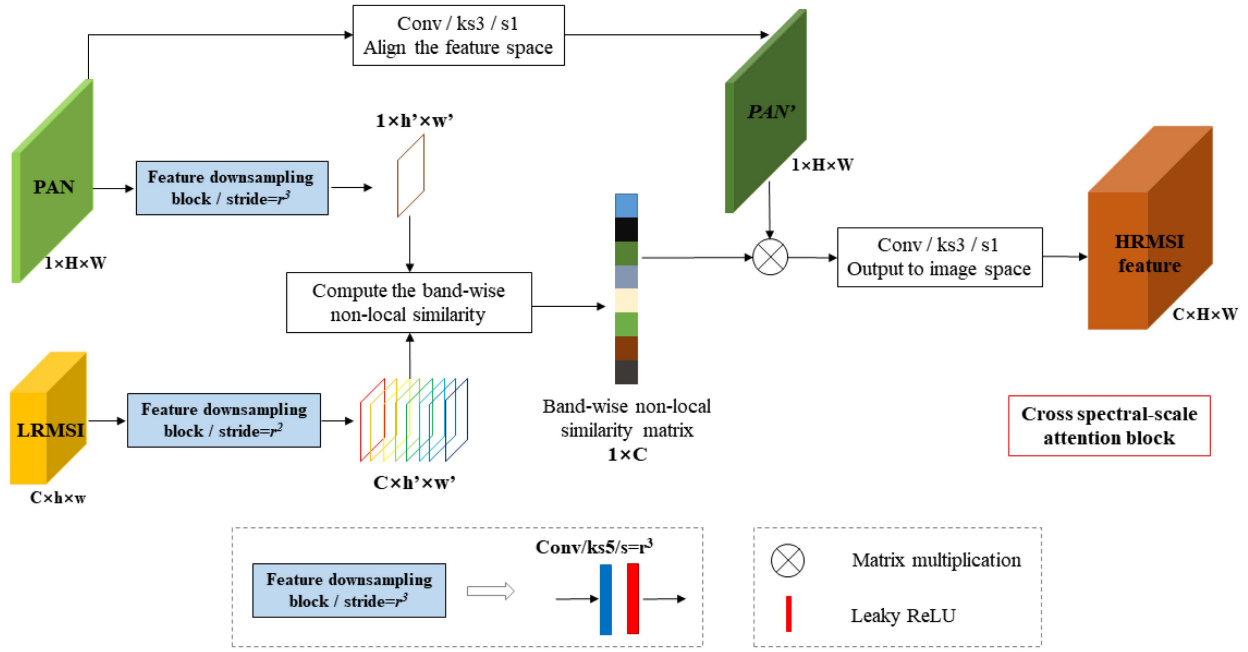


Fig. 3. Network structure of the proposed cross spectral-scale nonlocal attention block. The “Feature downsampling block” includes a convolution layer with kernel size 3 and a LeakyRelu activation function. The strides of convolution are set to r^2 and r^3 for the LRMSI and PAN images, respectively. The nonlocal similarity is computed by the bandwise dot-product operation. (“ r ” is the spatial resolution ratio between the LRMSI and PAN image.)

convolution operation and get $PAN'(P')$

$$(S_{1 \times C})^T \times P'_{1 \times H \times W} \Rightarrow \mathbb{M}'_{C \times H \times W} \quad (4)$$

where $\mathbb{M}'_{C \times H \times W}$ is the reconstructed HRMSI feature.

C. Cross Spatial-Scale Nonlocal Attention Block

The spatial resolution gap between the LRMSI and the PAN image is another major obstacle to achieving higher fusion performance. Some methods preupsample the LRMSI and concatenate it with the PAN image in the spectral dimension before entering it into the fusion model. This may introduce undesired spatial distortion or noise into the fused result. Therefore, to adapt to the spatial resolution gap between the LRHMI and PAN images and to benefit from both the local and nonlocal feature learning, we propose a cross spatial-scale nonlocal attention block, as shown in Fig. 4. This block could use the spatially cross-scale nonlocal similarity between the LRMSI and PAN images to reconstruct the HRMSI feature.

As shown in the left part of Fig. 4, we first measure the patchwise nonlocal similarity matrix between the encoded high-level features of LRMSI and PAN images. Then, we map this similarity to the low-level feature of the LRMSI image to obtain the low-level feature of the reconstructed HRMSI. Note that the patch-level nonlocal similarity could retain more local texture information compared to the pixelwise similarity. And it could explore the long-range dependency of the image feature. This combination of local and nonlocal feature learning could greatly improve the fused image quality.

Actually, the assumption of the similarity mapping theory is that the similarity between the images' high-level features

is closely related to the similarity between their low-level features. Generally, the high-level feature means the deep semantic information and the low-level feature represents the shallow spatial texture and spectral feature. For example, the trees in the same category (high-level feature) always have nearly the same spatial and spectral feature (low-level feature). Therefore, this computed spatially nonlocal similarity matrix could be shared between the high-level and low-level features

$$\psi(M_{i,j}^h, P_{m,n}^h) = \psi(M_{i,j}^l, \mathbb{M}_{m,n}^l) \quad (5)$$

where $M_{i,j}^h$ means the (i_{th}, j_{th}) unfolded patches of the LRMSI's high-level feature. $P_{m,n}^h$ means the (m_{th}, n_{th}) unfolded patches of the PAN image's high-level feature. And $\mathbb{M}_{m,n}^l$ means the (m_{th}, n_{th}) unfolded patches of the HRMSI's low-level feature. In fact, the theoretical basis of the similarity mapping is (5), which indicates that the computed spatially nonlocal similarity matrix could be shared between the high-level and low-level features. Equation (5) suggests that we could first extract the high-level feature from the input LRMSI and PAN images, then map the computed similarity between these two high-level features to the low-level feature of LRMSI and reconstruct the low-level feature of HRMSI.

In detail, we first obtain the high-level features of LRMSI and PAN images through the “high-level feature encoding block,” as shown in Fig. 4. The detailed network structure of the “high-level feature encoding block” is shown at the bottom of Fig. 4. It consists of two cascaded convolution blocks, which include the convolution layer, batch normalization, and LeakyRELU functions. Then, five “Resnet blocks” are used to extract the high-level features of the image. Then, we unfold the encoded

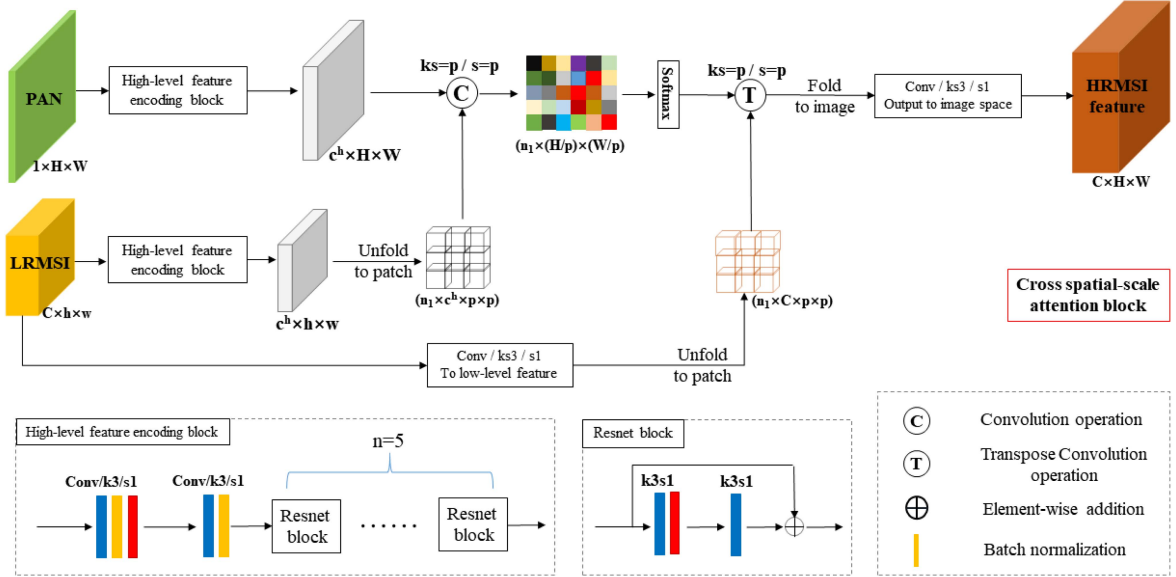


Fig. 4. Network structure of the proposed cross spatial-scale nonlocal attention block. Note that “p” is the unfolding patch size and “n₁” is the number of patches after the unfolding operation. “c^h” is the number of high-level feature bands. The “softmax” function is used to normalize the similarity matrix.

high-level feature of LRMSI into patches

$$M_{c^h \times h \times w}^h \rightarrow \text{unfold}() \rightarrow M_{n_1 \times c^h \times p \times p}^h \quad (6)$$

where $M_{c^h \times h \times w}^h$ is the encoded high-level feature of LRMSI. c^h is its band number (it is set to 32 in this study). n_1 is the number of patches after the $\text{unfold}()$ operation based on the patch size $p \times p$ ($n_1 = hw/(p \times p)$). And the unfolded patches are depicted by the “cube grid” in the middle part of Fig. 4.

To compute the cross-scale patch-level similarity with high efficiency, we use the convolution operation from the unfolded patches— $M_{n_1 \times c^h \times p \times p}^h$ (acting as the convolution kernel) onto the high-level feature of the PAN image (P^h)

$$f_{M_{n_1 \times c^h \times p \times p}^h}^h(P_{c^h \times H \times W}^h) \Rightarrow S_{n_1 \times (H/p) \times (W/p)} \quad (7)$$

where $P_{c^h \times H \times W}^h$ is the encoded high-level feature of the PAN image. $S_{n_1 \times (H/p) \times (W/p)}$ is the resulting nonlocal similarity matrix. Note that the stride of this convolution operation is set to the patch size- p . And the convolution operation is actually to compute the similarity between each of two patches of size $p \times p$. Then, as shown in Fig. 4, to reconstruct the HRMSI, the low-level feature of the LRMSI (M^l) is unfolded

$$M_{C \times h \times w}^l \rightarrow \text{unfold}() \rightarrow M_{n_1 \times C \times p \times p}^l \quad (8)$$

In the next step, as indicated by [39], we map the similarity matrix $S_{n_1 \times (H/p) \times (W/p)}$ to $M_{n_1 \times C \times p \times p}^l$ through the transpose convolution and get the reconstructed HRMSI feature as

$$f_{M_{n_1 \times C \times p \times p}^l}(S_{n_1 \times (H/p) \times (W/p)}) \Rightarrow \widehat{M}_{C \times H \times W}^l \quad (9)$$

where $f(\cdot)$ means to perform the transpose convolution operation using the kernel $M_{n_1 \times C \times p \times p}^l$ on the $S_{n_1 \times (H/p) \times (W/p)}$ with the stride of p . $\widehat{M}_{C \times H \times W}^l$ means the reconstructed low-level feature

of HRMSI. And then it is transferred to the original image space by a single convolution layer.

Note that the proposed cross spatial-scale attention block is inspired by [39], but there are two main differences between that study and our proposed network.

On the one hand, the cross-scale attention computed in [39] is between the original image and its downsampled version, while ours is computed between the LRMSI and PAN images on their original spatial resolutions. The downsampling operation in [39] may introduce errors and degrade the accuracy of the computed similarity matrix.

On the other hand, the nonlocal similarity in [39] is computed on low resolution scale and used on high resolution scale. While in our proposed model, the computation and use of the similarity matrix are on the same resolution scales. The resolution gap in [39] may degrade the resulting image quality.

D. Feature Aggregation Module

After getting the reconstructed HRMSI feature from the cross spatial-scale attention block, we concatenate it with the feature from the cross spectral-scale block. Then, as shown in Fig. 1, the “Feature aggregation module” including the “Dwconv”- with kernel sizes 11 and 1 is used to integrate the image features and output the fused HRMSI. Note that we add the residual connection [40] from the upsampled LRMSI (by the “bicubic”—interpolation method) to the final result to accelerate the training process.

E. Loss Function

In this part, different from the supervised methods, we design several loss functions to complete the training stage of the proposed unsupervised fusion network. The supervised methods could directly use the difference between the label and the fused

image to update the network parameters. While the unsupervised methods need to use the input images to constrain and improve the quality of the fused image.

1) *Bicubic Loss*: To guide the fused HRMSI to preserve the main spectral structure of the LRMSI, we adopt the “bicubic” upsampling consistency loss function

$$L_{bic_up} = \|M \uparrow - \widehat{M}\|_1 \quad (10)$$

where $M \uparrow$ is the upsampled LRMSI by the “bicubic” interpolation method. And \widehat{M} is the fused HRMSI.

2) *Detail Consistency Loss*: In fact, the spatial details involved in the HRMSI and PAN images are very similar. Therefore, we impose the detail consistency loss between the fused \widehat{M} and PAN image (P) by their high-frequency features

$$L_{dc} = r(\widehat{M} - M \uparrow, P - P_b) \quad (11)$$

where P_b is the blurred PAN image through the Gaussian kernel. And $r(x, y)$ means the calculation of the correlation coefficient between x and y

$$r(x, y) = \frac{E((x - E(x))(y - E(y)))}{std(x)std(y)} \quad (12)$$

where $E()$ means the calculation of the expected value and $std()$ is the standard deviation calculation function.

3) *Local Spectral and Spatial Consistency Loss*: The local texture similarity losses in both spectral and spatial domains are added by the SSIM index [41]

$$L_{SSIM_spe} = 1 - SSIM(Down_{spa}(\widehat{M}), M) \quad (13)$$

$$L_{SSIM_spa} = 1 - SSIM(Down_{spe}(\widehat{M}), P) \quad (14)$$

$$L_{SSIM} = L_{SSIM_spe} + L_{SSIM_spa} \quad (15)$$

where $Down_{spa}$ means the spatial downsampling operation through the Gaussian kernel. And $Down_{spe}$ is the spectral downsampling operation by the average function.

4) *Total Loss Function*: The total loss function is the combination of the above three loss functions as

$$L_{total} = \alpha_1 L_{bic_up} + \alpha_2 L_{dc} + \alpha_3 L_{SSIM} \quad (16)$$

where α_1 , α_2 , and α_3 are the weights of the corresponding loss term and they are set to 10, 10, and 50, respectively.

IV. EXPERIMENTAL RESULTS

In this section, we take the fusion experiments and show the fused results. First, we describe the two datasets used in this study and the training details. Then, we list the comparison methods and quality measure metrics. Finally, we show the qualitative and quantitative fusion results of different methods on two datasets on reduced and full resolution scales.

A. Datasets

We select two datasets including WorldView2 (WV2) and GaoFen2 (GF2) satellite images to verify the superiority of the proposed fusion method. The spatial resolutions of these two datasets are 0.5 and 0.8 m for PAN image, 2.0 and 3.2 m for

LRMSI, respectively. And their number of bands are 8 and 4 in fusion experiments. The spatial sizes of these two datasets are 5059×2145 and 6907×7300 for the LRMSI, 20236×8580 and 27628×29200 for the PAN image.

The fusion experiments are conducted on reduced and full resolution datasets. The reduced resolution dataset is generated according to the Wald protocol [42] by downsampling the original LRMSI and PAN image with the Gaussian kernel. The original LRMSI is then regarded as the reference image. The patch size is set to 64 and 256 for the LRMSI and PAN images, respectively. And we randomly select 90% of the cropped patches in the training stage and the last is for the performance test. Note that the reduced resolution experiments are actually to compare the different methods on the simulated dataset, which lacks practicality. And the full resolution experiments on the original dataset are more practical in real-world applications.

B. Training Details

Due to the different number of training image pairs on the reduced and full resolution scales, we set the training epochs to 500 and 50 for these two scales’ datasets, respectively. The batch size is set to 10. And the learning rate is initialized to $1e-4$, and decays with a rate of 0.1 in half of the training epochs. The ADAM optimizer is selected to update the model parameters with the β_1 of 0.9 and β_2 of 0.999. The feature bands are all set to 32 except the one specified. All the experiments are run under the paddle 2.4.0 framework on a single V100 GPU with 32 GB memory.

C. Comparison Methods and Quality Measures Metrics

We compare the proposed CSFNet with several SOTA pansharpening methods, including IHS [8], SFIM [11], Wavelet [12], MTF_GLP [23], and MTF_GLP_HPM [23], which belong to the traditional methods. As for the DL-based methods, we select several unsupervised and supervised methods. The first class includes LDPnet [43] and PanGAN [21]. Like the proposed CSFNet, these two methods are trained on reduced and full resolution datasets, respectively. For the supervised methods, Pannet [17], TFNet [44], and PanFormer [45] are selected to compare the fusion performance. For these three methods, due to the requirement of the reference image to supervise the training process, we only train them on the reduced resolution dataset. And the full resolution’s performance test is done by the trained parameters on the reduced dataset.

The quality indices used to measure and compare the performance of different methods on the reduced resolution datasets include Spectral Angle Mapper (SAM) [46], relative dimensionless global error in synthesis (ERGAS) [47], universal image quality index (UIQI) [48], and root-mean-squared error (rmse). SAM measures spectral distortion in fused images. ERGAS, UIQI, and rmse evaluate the spectral and spatial quality of the fused results comprehensively. For the full resolution experiments, due to the lack of reference image, we choose the spectral index D_λ [49], spatial distortion index D_s [49] and QNR [49]

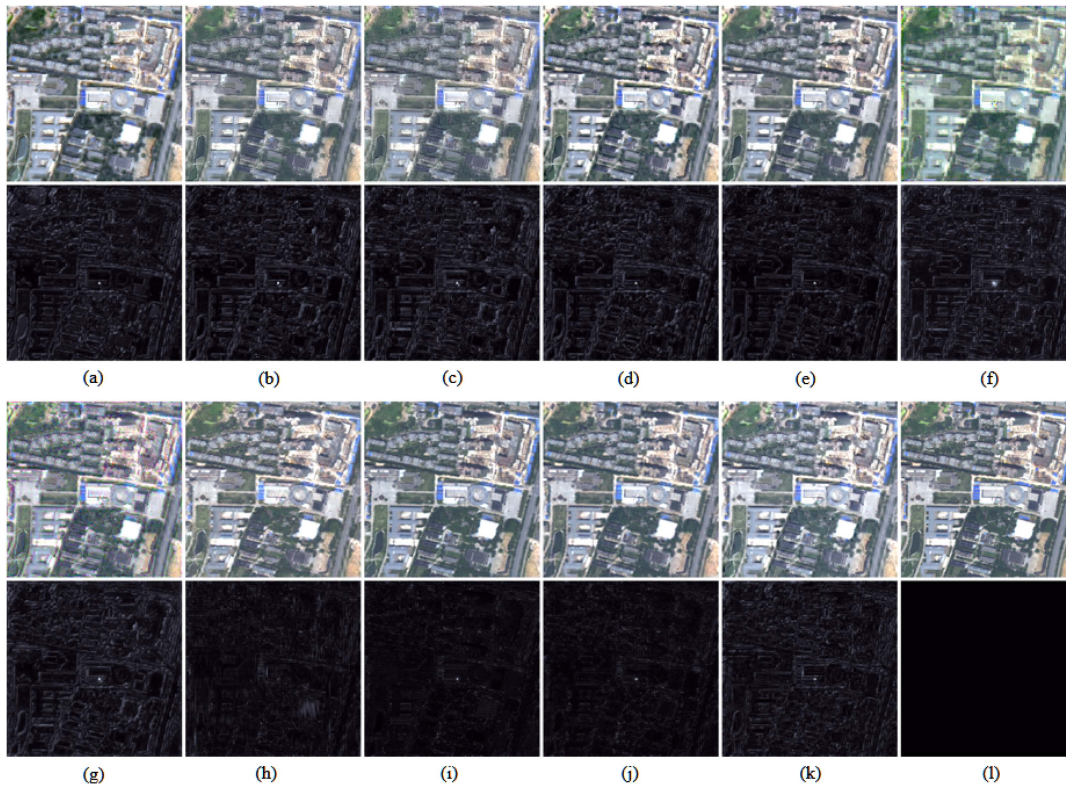


Fig. 5. Fused results (odd row) and error maps (even row) of different methods on the reduced resolution WV2 dataset. (a) IHS. (b) SFIM. (c) Wavelet. (d) MTF_GLP. (e) MTF_GLP_HPM. (f) LDPnet. (g) PanGAN. (h) Pannet. (i) TFNet. (j) PanFormer (k) Ours. (l) Ground truth.

to compare the spectral and spatial fidelity of different methods' results. The generalized QNR [50] (GQNR) has also been proposed to compare the different methods. But considering the similarity between the GQNR and QNR indices, we only use the QNR index to compare the different methods. Note that the supervised methods are highlighted with dashed lines in the quantitative results. And the red font in the quantitative results means the best result among all unsupervised methods except the supervised methods, while the bold font means the best result among all methods, and the underlined results represent the second rank among all methods.

D. Fusion Results on Reduced Resolution Dataset

We first conduct experiments on the reduced WV2 and GF2 datasets to compare the results of different methods visually and quantitatively. Note that these experiments are conducted on the simulated datasets, which reduces their significance in real-world applications and lack of practicality.

The fused results of these two simulated datasets are shown in Figs. 5 and 6, respectively. We show both the fused results and the error maps to clearly compare the different methods. Note that the error map represents the average of all bands' absolute difference between the reference and fused result. For the WV2 dataset, as shown in Fig. 5(k), our results show the best detail restoration and spectral fidelity among the unsupervised methods. Compared to the supervised methods, the unsupervised methods have no reference image to guide the training

process. So they perform slightly worse than the supervised methods on the simulated datasets. The fusion results of SFIM and Wavelet all suffer from the blurring effect, especially at the edge of buildings and roads, as shown in the Fig. 5(b) and (c). And severe spectral distortion occurs on the fused result of the LDPnet method. The quantitative results in Table I also show that our method achieves the best rank in three indices except for the supervised methods, as indicated by the red font in this table. And LDPnet gets the worst rank in four indices, which is consistent with the visual result in Fig. 5(f). Note that the best result in each column is highlighted in the bold style and underlined for the inferior result. And the red font indicates the best rank among the unsupervised methods.

We also perform experiments on the GF2 dataset with different land covers to demonstrate the robustness of our method. The qualitative results in Fig. 6 show that our method achieves the best spatial and spectral fidelity compared to others (except the supervised method). Note that the PanGAN method suffers from the spatial artifact, as shown in Figs. 5(g) and 6(g). The reason may be that the spectral and spatial constraints used in the loss functions of this method are not appropriate totally.

E. Fusion Results on Full Resolution Datasets

In this part, we conduct the fusion experiments on the original full resolution dataset, which is more practical in real applications than the experiments on the reduced resolution dataset.

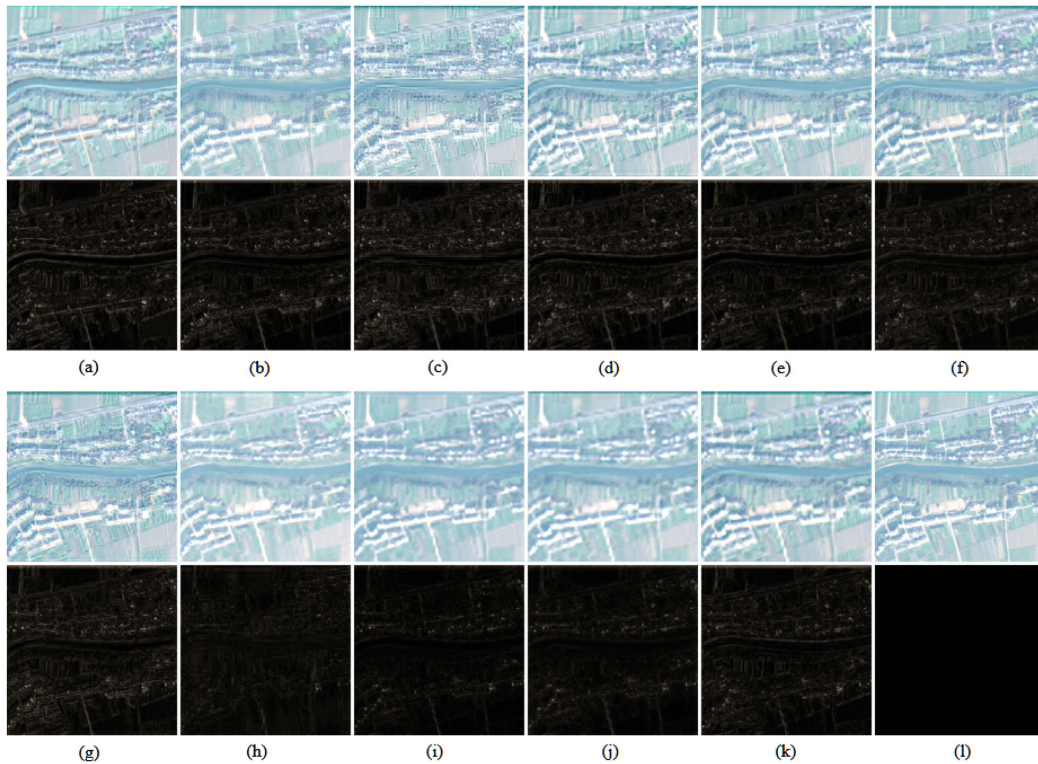


Fig. 6. Fused results (odd row) and error maps (even row) of different methods on the reduced resolution GF2 dataset. (a) IHS. (b) SFIM. (c) Wavelet. (d) MTF_GLP. (e) MTF_GLP_HPM. (f) LDPnet. (g) PanGAN. (h) Pannet. (i) TFNet. (j) PanFormer (k) Ours. (l) Ground truth.

TABLE I
QUANTITATIVE RESULT ON WV2 DATASET

Method	SAM(\downarrow)	ERGAS(\downarrow)	UIQI(\uparrow)	RMSE(\downarrow)	D_λ (\downarrow)	D_s (\downarrow)	QNR(\uparrow)
IHS	0.088	4.957	0.436	0.029	<u>0.074</u>	0.141	0.797
SFIM	0.083	5.525	0.394	0.033	0.116	0.133	0.772
Wavelet	0.081	5.359	0.380	0.032	0.108	0.135	0.778
MTF_GLP	0.087	4.837	0.459	0.030	0.145	0.133	0.746
MTF_GLP_HPM	0.083	5.027	0.461	0.030	0.146	0.135	0.744
LDPnet	0.126	6.732	0.339	0.036	0.132	0.097	0.785
PanGAN	0.100	5.234	0.405	0.030	0.117	0.143	0.763
Pannet	0.068	3.232	0.527	0.019	0.170	0.120	0.727
TFNet	0.051	2.482	0.589	0.016	0.104	0.074	<u>0.830</u>
PanFormer	0.061	2.770	0.533	0.017	0.133	0.081	0.798
Ours	0.081	5.134	0.472	0.028	0.073	<u>0.076</u>	0.843

The qualitative fusion results on WV2 and GF2 datasets are shown in the Figs. 7 and 8, respectively. It can be seen that our results show better spectral and spatial fidelity than other methods. For the WV2 dataset, as shown in Fig. 7, some methods' fusion results suffer from the spatial blurring effect, such as SFIM, Wavelet and LDPnet. And several methods suffer from spectral artifacts and distortion, such as LDPnet and PanGAN, as shown in Figs. 7(g) and 8(g), (h). Note that the supervised methods are trained on the simulated reduced resolution dataset, so they perform poorly on the full resolution images in restoring the

spatial details. This could be obviously seen from their visual results, especially the Pannet method.

The quantitative results of the nonreference quality indices also show the superiority of our methods over others, as shown in the last three columns of Tables I and II. Our method achieves the best results on all indices except the D_s index on the GF2 dataset, as shown in Table II. And even the MTF_GLP achieves the best rank on the D_s index, which advances our method by 0.003, it is hard to achieve a satisfactory spectral fidelity, as indicated by its D_λ index's result. All in all, our method achieves the best



Fig. 7. Fused results of different methods on the full resolution WV2 dataset. (a) Upsampled LRMSI. (b) IHS. (c) SFIM. (d) Wavelet. (e) MTF_GLP. (f) MTF_GLP_HPM. (g) LDPnet. (h) PanGAN. (i) Pannet. (j) TFNet. (k) PanFormer. (l) Ours.

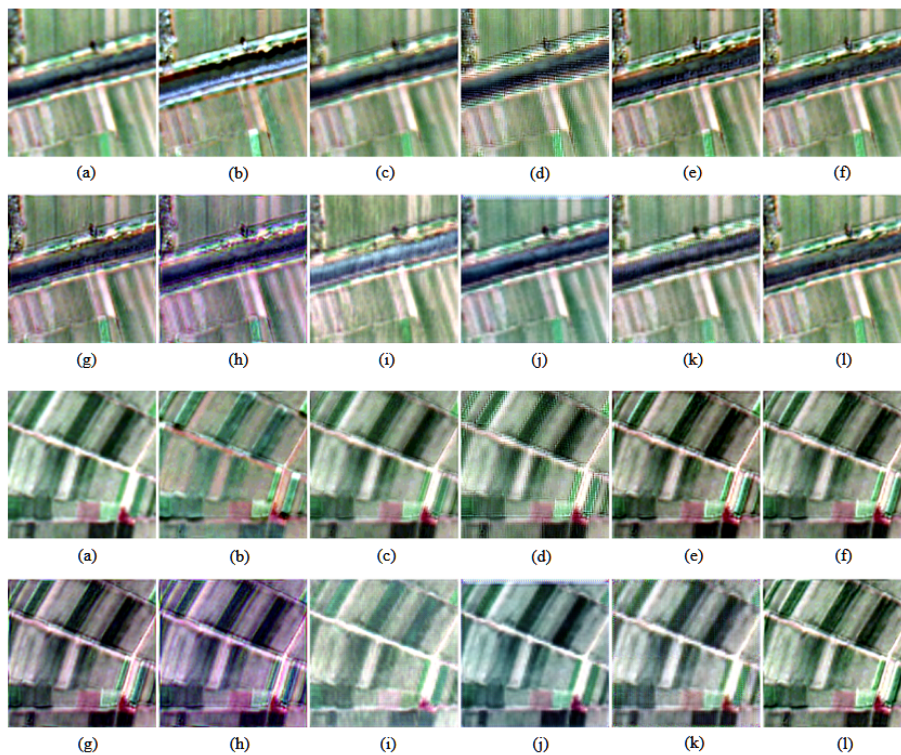


Fig. 8. Fused results of different methods on the full resolution GF2 dataset. (a) Upsampled LRMSI. (b) IHS. (c) SFIM. (d) Wavelet. (e) MTF_GLP. (f) MTF_GLP_HPM. (g) LDPnet. (h) PanGAN. (i) Pannet. (j) TFNet. (k) PanFormer. (l) Ours.

TABLE II
QUANTITATIVE RESULT ON GF2 DATASET

Method	SAM(\downarrow)	ERGAS(\downarrow)	UIQI(\uparrow)	RMSE(\downarrow)	D_λ (\downarrow)	D_s (\downarrow)	QNR(\uparrow)
IHS	0.045	2.840	0.243	0.029	0.120	0.373	0.550
SFIM	0.046	2.844	0.307	0.028	0.096	0.169	0.753
Wavelet	0.044	2.941	0.258	0.029	<u>0.082</u>	0.169	0.764
MTF_GLP	0.047	2.879	0.293	0.029	0.102	0.142	<u>0.774</u>
MTF_GLP_HPM	0.046	2.858	0.296	0.028	0.104	<u>0.142</u>	0.772
LDPnet	0.045	2.853	0.299	0.028	0.098	0.151	0.761
PanGAN	0.052	3.212	0.319	0.032	0.119	0.170	0.733
Pannet	0.031	1.798	0.626	0.018	0.116	0.238	0.678
TFNet	0.029	1.630	0.641	0.016	0.084	0.182	0.752
PanFormer	0.034	1.644	0.552	0.017	0.118	0.185	0.723
Ours	0.037	2.131	0.451	0.021	0.081	0.145	0.783

TABLE III
AVERAGE QUANTITATIVE RESULTS ON FULL GF2 DATASET WITH DIFFERENT STRUCTURES; “W/O CROSS SPECTRAL-SCALE” MEANS WITHOUT CROSS SPECTRAL-SCALE BLOCK; “BASE LINE” MEANS THAT WITH BOTH BLOCKS

encoder/decoder	D_λ (\downarrow)	D_s (\downarrow)	QNR(\uparrow)
w/o two cross-scale blocks	0.148	0.152	0.731
w/o cross spectral-scale	0.074	0.188	0.752
w/o cross spatial-scale	0.082	0.154	0.772
base line	0.081	0.145	0.783

balance between spatial and spectral fidelity on two datasets, which demonstrates the superiority of our method.

V. DISCUSSION

This section conducts the ablation study to demonstrate the effectiveness of the designed network structures and loss functions. Then, we test the different unfolding kernel sizes to achieve the best balance between local and nonlocal feature learning. Finally, we compare the computational complexity of all the methods.

A. Network Structures

To demonstrate the effectiveness of the proposed network modules, we gradually remove the cross spectral-scale and cross spatial-scale nonlocal attention blocks to measure their contributions, respectively. As shown in Table III, which are experimental results on the full resolution GF2 dataset, it could be inferred that two designed cross-scale blocks all improve the fused image quality from the first three rows of this table. And the cross spectral-scale block contributes more compared to the cross spatial-scale block. For example, the D_s and QNR indices improve from 0.188 and 0.752 to 0.154 and 0.772. The increments of percentage are 18.09% and 2.66%, respectively. Finally, the combination of these two blocks further improves the fusion performance, as shown in the last row of Table III.

The quantitative results on the reduced GF2 dataset also demonstrate the effectiveness of two cross-scale blocks. As shown in Table IV, the quality indices have all improved after

TABLE IV
AVERAGE QUANTITATIVE RESULTS ON REDUCED GF2 DATASET WITH DIFFERENT STRUCTURES

encoder/decoder	SAM(\downarrow)	ERGAS(\downarrow)	UIQI(\uparrow)	RMSE(\downarrow)
w/o two cross-scale blocks	0.055	2.541	0.417	0.025
w/o cross spectral-scale	0.038	2.191	0.433	0.022
w/o cross spatial-scale	0.042	2.235	0.429	0.022
base line	0.037	2.131	0.451	0.021

TABLE V
DIFFERENT LOSS FUNCTION SETTINGS (“UPSAMPLE” OR “DOWNSAMPLE” CONSISTANCY) AND THEIR QUANTITATIVE RESULTS ON FULL GF2 DATASET

method	D_λ (\downarrow)	D_s (\downarrow)	QNR(\uparrow)
upsample	0.081	0.145	0.783
downsample	0.081	0.157	0.776

the addition of each cross-scale block. All in all, the above experimental results verify the effectiveness of these two designed network modules.

B. Loss Functions

We conduct experiments on loss functions to demonstrate the necessity of each of them. First, we test the “bicubic” upsampling consistency loss function. This loss function could preserve the original spectral structure of the LRMSI well. And we compare this loss function with the “bicubic” downsampling loss function, which means that the fused HRMSI is downsampled and should be consistent with the input LRMSI as much as possible. As shown in Tables V and VI, the fusion results on both high and low resolution GF2 datasets demonstrate the effectiveness of the ‘bicubic’ upsampling consistency loss function.

Then, as shown in Tables VII and VIII, we, respectively, remove three proposed loss functions to verify their effectiveness. It could be inferred that these loss functions all contribute to the final excellent fusion performance, especially the L_{bic_up}

TABLE VI
DIFFERENT LOSS FUNCTION SETTINGS (“UPSAMPLE” OR “DOWNSAMPLE” CONSISTENCY) AND THEIR QUANTITATIVE RESULTS ON REDUCED GF2 DATASET

method	SAM(↓)	ERGAS(↓)	UIQI(↑)	RMSE(↓)
upsample	0.037	2.131	0.451	0.021
downsample	0.037	2.210	0.429	0.022

TABLE VII
DIFFERENT LOSS FUNCTION COMBINATIONS AND ITS QUANTITATIVE RESULTS ON FULL GF2 DATASET

L_{bic_up}	L_{dc}	L_{SSIM}	D_λ (↓)	D_s (↓)	QNR(↑)
✗	✓	✓	0.080	0.220	0.719
✓	✗	✓	0.098	0.169	0.752
✓	✓	✗	0.202	0.165	0.670
✓	✓	✓	0.081	0.145	0.783

TABLE VIII
DIFFERENT LOSS FUNCTION COMBINATIONS AND ITS QUANTITATIVE RESULTS ON REDUCED GF2 DATASET

L_{bic_up}	L_{dc}	L_{SSIM}	SAM(↓)	ERGAS(↓)	UIQI(↑)	RMSE(↓)
✗	✓	✓	0.054	4.233	0.423	0.043
✓	✗	✓	0.037	2.197	0.470	0.022
✓	✓	✗	0.041	2.799	0.305	0.028
✓	✓	✓	0.037	2.131	0.451	0.021

TABLE IX
AVERAGE QUANTITATIVE RESULT ON FULL GF2 DATASET WITH DIFFERENT UNFOLDING PATCH SIZE IN THE CROSS-SPATIAL SCALE BLOCK

unfolding patch size	D_λ (↓)	D_s (↓)	QNR(↑)
3	0.081	0.145	0.783
5	0.079	0.286	0.661
7	0.082	0.153	0.776

TABLE X
AVERAGE QUANTITATIVE RESULT ON REDUCED GF2 DATASET WITH DIFFERENT UNFOLDING PATCH SIZE IN THE CROSS-SPATIAL SCALE BLOCK

unfolding patch size	SAM(↓)	ERGAS(↓)	UIQI(↑)	RMSE(↓)
3	0.037	2.131	0.451	0.021
5	0.037	2.143	0.436	0.021
7	0.037	2.193	0.467	0.022

and L_{SSIM} which make significant contributions. Note that the L_{SSIM} mainly improves the spectral index- D_λ from 0.202 to 0.081 (improved by 59.90%), as shown in the last two rows of Table VII.

C. Unfolding Kernel Size

To take a balance between the local and nonlocal feature learning ability in computing the spatially nonlocal similarity matrix, we take the experiment on different unfolding patch sizes- p in cross spatial-scale attention block. As shown in Tables IX and X, kernel size 3 achieves the best fusion performance, which means it could extract the global and similar local features effectively.

TABLE XI
AVERAGE TESTING TIME AND MODEL PARAMETERS ON REDUCED WV2 DATASET

Method	Time(s)(↓)	Parameters(m)(↓)
IHS	0.139	—
SFIM	0.062	—
Wavelet	0.021	—
MTF_GLP	0.411	—
MTF_GLP_HPM	0.131	—
LDPnet	0.064	6.417
PanGAN	0.003	0.916
Pannet	0.004	0.084
TFNet	0.013	2.204
PanFormer	0.031	1.533
Ours	0.027	0.259

And it could achieve the best balance between learning the local texture feature and nonlocal feature similarity.

D. Effect of the Cross Scale Blocks

To verify the effect of the designed two cross scale blocks, we visualize the learned feature by the cross spectral-scale and cross spatial-scale blocks, as shown in Fig. 9. It could be clearly seen that the cross spectral-scale block mainly reconstructs the spectral feature, while the cross spatial-scale block mainly reconstructs the spatial feature. Therefore, these visualization results verify the effectiveness of the proposed cross scale blocks.

E. Comparison of Computational Complexity

In this part, we list the inference time and the number of parameters of all methods to compare their fusion efficiency. This calculation is performed on the WV2 dataset with the patch size of $8 \times 64 \times 64$ (LRMSI) and $1 \times 256 \times 256$ (PAN image).

As shown in Table XI, the traditional methods all suffer from the large inference time consumption, especially the IHS, MTF_GLP and MTF_GLP_HPM. In comparison, DL-based methods all have an immediate inference time. This is because the DL-based methods could benefit from GPU acceleration, which is the unique characteristic of DL-based methods. And the traditional methods are mostly designed to compute in the normal device. Even though the supervised methods could achieve great fusion performance on reduced simulation datasets, their fusion performance on full resolution dataset is worse than others. Furthermore, among the unsupervised methods, although the PanGAN method costs less inference time, its fusion performance is inferior to ours, as shown in section IV-D and IV-E. And the increment of the inference time of our method compared with the PanGAN is 0.024 s, which is negligible. Therefore, it can be concluded that our method achieves the best balance between the fusion performance and model complexity.

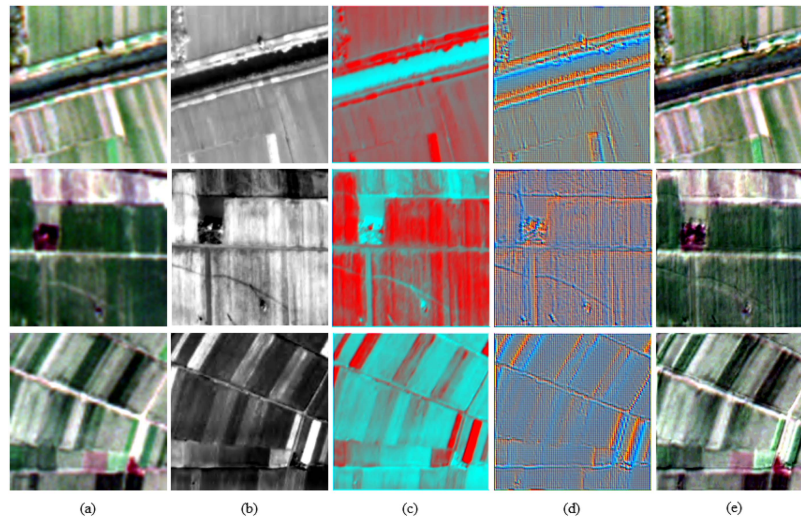


Fig. 9. Visualization of the learned feature by two cross scale blocks on the full resolution GF2 dataset. (a) Upsampled LRMSI. (b) PAN image. (c) Learned feature by cross spectral-scale block. (d) Learned feature by cross spatial-scale block. (e) Fusion results.

VI. CONCLUSION

In this article, in order to correctly learn and utilize the spectral and spatial correlation between LRMSI, HRMSI and PAN images to reconstruct the fusion result, we design two cross spectral-scale and cross spatial-scale nonlocal attention blocks. The designed cross spectral-scale block computes the bandwise nonlocal similarity on low-resolution images and maps this similarity to the high-resolution scale to reconstruct the HRMSI feature. The designed cross spatial-scale block computes the patch-level nonlocal similarity on the high-level feature of the input images. Then, it maps this similarity to the original LRMSI image feature to reconstruct the HRMSI feature. So the proposed cross spatial-scale block could effectively combine the advantages of nonlocal and local feature learning. Finally, a “Feature aggregation module” integrates the HRMSI features constructed by these two blocks and outputs the fused HRMSI. Experimental results on two reduced and full resolution datasets demonstrate the effectiveness and superiority of the proposed fusion network.

ACKNOWLEDGMENT

The authors would like to thank the Key Laboratory of Natural Resources Monitoring in Tropical and Subtropical Area of South China, Ministry of Natural Resources. The authors would also like to thank Baidu AI Studio for providing the computing power supports.

REFERENCES

- [1] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, and J. Tian, “Robust feature matching for remote sensing image registration via locally linear transforming,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 12, pp. 6469–6481, Dec. 2015.
- [2] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, “Deep learning-based fusion of landsat-8 and sentinel-2 images for a harmonized surface reflectance product,” *Remote Sens. Environ.*, vol. 235, 2019, Art. no. 111425.
- [3] 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.
- [4] G. Vivone et al., “A critical comparison among pansharpening algorithms,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.
- [5] G. Vivone et al., “A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021.
- [6] X. Meng et al., “A large-scale benchmark data set for evaluating pansharpening performance: Overview and implementation,” *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 18–52, Mar. 2021.
- [7] G. Vivone et al., “A critical comparison of pansharpening algorithms,” in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 191–194.
- [8] T.-M. Tu, S.-C. Su, H.-C. Shyu, and P. S. Huang, “A new look at IHS-like image fusion methods,” *Inf. Fusion*, vol. 2, no. 3, pp. 177–186, 2001.
- [9] “Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening,” U.S. Patent 6 011 875, Jan. 4, 2000.
- [10] B. Aiazzi, S. Baronti, and M. Selva, “Improving component substitution pansharpening through multivariate regression of MS +Pan data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.
- [11] J. G. 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.
- [12] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, “Multiresolution-based image fusion with additive wavelet decomposition,” *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.
- [13] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, “Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis,” *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.
- [14] “MTF-tailored multiscale fusion of high-resolution ms and pan imagery,” *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.
- [15] N. Yokoya, T. Yairi, and A. Iwasaki, “Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion,” *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 528–537, Feb. 2012.
- [16] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, “Pansharpening by convolutional neural networks,” *Remote Sens.*, vol. 8, no. 7, 2016, Art. no. 594.
- [17] 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.
- [18] S. Li, Y. Tian, C. Wang, H. Wu, and S. Zheng, “Hyperspectral image super-resolution network based on cross-scale nonlocal attention,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–15, 2023.

- [19] C. Kwan, B. Budavari, M. Dao, B. Ayhan, and J. F. Bell, "Pansharpening of mastcam images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2017, pp. 5117–5120.
- [20] S. Li, Y. Tian, H. Xia, and Q. Liu, "Unmixing-based PAN-guided fusion network for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2022.
- [21] 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.
- [22] H. Zhou, Q. Liu, and Y. Wang, "Unsupervised cycle-consistent generative adversarial networks for pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5408814, doi: [10.1109/TGRS.2022.3166528](https://doi.org/10.1109/TGRS.2022.3166528).
- [23] Y. Qu, R. K. Baghbaderani, H. Qi, and C. Kwan, "Unsupervised pansharpening based on self-attention mechanism," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 4, pp. 3192–3208, Apr. 2021.
- [24] A. Vaswani et al., "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: <http://arxiv.org/abs/1706.03762>
- [25] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5791–5800.
- [26] Z. Lu, J. Li, H. Liu, C. Huang, L. Zhang, and T. Zeng, "Transformer for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 457–466.
- [27] F. Zhang, K. Zhang, and J. Sun, "Multiscale spatial-spectral interaction transformer for PAN-sharpening," *Remote Sens.*, vol. 14, no. 7, 2022, Art. no. 1736.
- [28] W. G. C. Bandara and V. M. Patel, "Hypertransformer: A textural and spectral feature fusion transformer for pansharpening," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1767–1777.
- [29] C. Zhou, J. Zhang, J. Liu, C. Zhang, R. Fei, and S. Xu, "PercepPan: Towards unsupervised pan-sharpening based on perceptual loss," *Remote Sens.*, vol. 12, no. 14, 2020, Art. no. 2318.
- [30] S. Seo et al., "UPSNet: Unsupervised pan-sharpening network with registration learning between panchromatic and multi-spectral images," *IEEE Access*, vol. 8, pp. 201199–201217, 2020.
- [31] W. Diao, F. Zhang, J. Sun, Y. Xing, K. Zhang, and L. Bruzzone, "ZeR-GAN: Zero-reference GAN for fusion of multispectral and panchromatic images," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published, doi: [10.1109/TNNLS.2021.3137373](https://doi.org/10.1109/TNNLS.2021.3137373).
- [32] Q. Xu, Y. Li, J. Nie, Q. Liu, and M. Guo, "UPanGAN: Unsupervised pansharpening based on the spectral and spatial loss constrained generative adversarial network," *Inf. Fusion*, vol. 91, pp. 31–46, 2023.
- [33] D. Glasner, S. Bagon, and M. Irani, "Super-resolution from a single image," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 349–356.
- [34] M. Zontak and M. Irani, "Internal statistics of a single natural image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 977–984.
- [35] G. Freedman and R. Fattal, "Image and video upscaling from local self-examples," *ACM Trans. Graph.*, vol. 30, no. 2, Apr. 2011.
- [36] D. Liu et al., "Non-local recurrent network for image restoration," in *Proc. 32th Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1680–1689.
- [37] T. Dai, J. Cai, Y. Zhang, S.-T. Xia, and L. Zhang, "Second-order attention network for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11057–11066.
- [38] S. Luo, S. Zhou, Y. Feng, and J. Xie, "Pansharpening via unsupervised convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4295–4310, 2020.
- [39] Y. Mei, Y. Fan, Y. Zhou, L. Huang, T. S. Huang, and H. Shi, "Image super-resolution with cross-scale non-local attention and exhaustive self-exemplars mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5690–5699.
- [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] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity," vol. 13, no. 4, pp. 600–612, 2003.
- [42] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
- [43] J. Ni et al., "LDP-Net: An unsupervised pansharpening network based on learnable degradation processes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5468–5479, 2022.
- [44] X. Liu, Q. Liu, and Y. Wang, "Remote sensing image fusion based on two-stream fusion network," *Inf. Fusion*, vol. 55, pp. 1–15, 2020.
- [45] H. Zhou, Q. Liu, and Y. Wang, "PanFormer: A transformer based model for pan-sharpening," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2022, pp. 1–6.
- [46] L. Loncan et al., "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.
- [47] 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.
- [48] Z. Wang and A. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.
- [49] "Multispectral and panchromatic data fusion assessment without reference," *Photogrammetric Eng. Remote Sens.*, vol. 74, no. 2, 2008.
- [50] C. Kwan, B. Budavari, A. C. Bovik, and G. Marchisio, "Blind quality assessment of fused worldview-3 images by using the combinations of pansharpening and hypersharpening paradigms," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1835–1839, Oct. 2017.



Shuangliang Li received the B.S. degree in geographical science from Hubei University, Wuhan, China, in 2019. He is currently working toward the M.S. degree in photogrammetry and remote sensing from China University of Geosciences, Wuhan, China.

His research interests include remote sensing image fusion and deep learning.



Yugang Tian (Member, IEEE) received the B.E. degree in surveying and mapping engineering from Wuhan University, Wuhan, China, in 2000, the M.S. degree in geodesy and geomatics engineering from Wuhan University, in 2003, and the Ph.D. degree in physical geography from Beijing Normal University, Beijing, China, in 2006.

Since 2009, he has been an Associate Professor with the School of Geography and Information Engineering, China University of Geosciences, Wuhan, China. From 2014 to 2015, he was a Visiting Scholar

with the Department of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA. His research interests include urban and environmental monitoring, image processing, and pattern cognition.



Cheng Wang received the B.E. degree in remote sensing science and technology from China University of Geosciences, Wuhan, China, in 2021, where he is currently working toward the M.S. degree in photogrammetry and remote sensing.

His research interests include remote sensing image analysis and semantic segmentation.



Hongxian Wu received the B.E. degree in remote sensing science and technology and the M.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011 and 2013, respectively.

He is currently working with Surveying and Mapping Institute, Lands and Resource Department of Guangdong Province, Guangzhou, China. His research interests include remote sensing image processing and application, protection and utilization of cultivated land resources.



Shaolan Zheng received the B.E. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2011, and the M.S. degree in surveying and mapping engineering from Wuhan University, in 2013.

She is currently working with Surveying and Mapping Institute, Lands and Resource Department of Guangdong Province, Guangzhou, China. Her research interests include remote sensing image processing and application, protection and utilization of cultivated land resources.