IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 October 2023, accepted 21 October 2023, date of publication 25 October 2023, date of current version 3 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3327427

RESEARCH ARTICLE

CGSNet: Channel Group Shuffling Network for Remote Sensing Image Fusion

HONGHUI JIANG[1], (Member, IEEE), HU PENG[2], (Member, IEEE), AND GUOZHENG ZHANG[3], (Member, IEEE)

[1]Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241003, China
[2]School of Instrument Science and Opto-Electronics Engineering, Hefei University of Technology, Hefei 230009, China
[3]School of Mechanical Engineering, Anhui Technical College of Mechanical and Electrical Engineering, Wuhu 241003, China

Corresponding author: Guozheng Zhang (jenuel@163.com)

This work was supported in part by the Academic Funding Project for Top-Notch Talents in Disciplines (Specialties) of Colleges and Universities under Grant gxbjZD2020108; and in part by the 2023 Anhui Province Higher Education Research Project, under Grant 2023AH052692.

ABSTRACT High-resolution multi-spectral (HRMS) images have been widely used in various fields, however, they can not be directly obtained due to the physical hardware limits of existing remote sensing satellites. Therefore, the pansharpening technique has been widely explored as an effective tool to generate HRMS images by fusing the complementary information of low-resolution multi-spectral (LRMS) images and high-resolution panchromatic (PAN) images. Existing deep learning-based pansharpening methods mainly focus on enhancing the spatial representation ability of the network, while paying little attention to modeling spectral dependencies in spite of its significance for remote sensing data. In this paper, we propose a simple yet effective channel group shuffling (CGS) module to explore the implicit relationships with regard to the adjacent and cross-channels while considering spatial information. To be specific, the proposed CGS module consists of two components: the channel group module and the feature shuffle fusion module. The former enhances the diversity of spectral information and cross-channel information communications while ensuring the spectral order of the input feature. The latter integrates the cross-group feature maps with rich spatial-spectral information. Equipped with the proposed functional module, our image fusion network, dubbed CGSNet, produces favorable results against existing state-of-the-art counterparts over various satellite datasets. Ablation studies further verify the flexibility and effectiveness of our core design.

INDEX TERMS Pansharpening, deep learning, channel group, feature shuffle fusion.

I. INTRODUCTION

The rapid development of remote sensing satellites facilitates the acquisition of remote sensing data, which thus receives much attention from image processing and remote sensing communities. Due to the physical limitations, however, existing remote sensing satellite sensors commonly capture both low-resolution multi-spectral (LRMS) images containing abundant spectral information and high-resolution panchromatic (PAN) images showing rich spatial details, instead of directly observing high-resolution multi-spectral (HRMS) images that are desirable in remote sensing applications, such as object detection [1], change detection [2], [3], environmental monitoring [4], [5], and mapping services [6], [7], [8]. Therefore, the pansharpening technique is developed to produce HRMS images by fusing the complementary information from the obtained MS images and PAN images. Over the past decades, pansharpening has been widely explored and made prominent achievements.

Traditional pansharpening approaches can be roughly divided into three streams, including component substitution (CS) methods [9], [10], multiresolution analysis (MRA) approaches [11], [12], and variational optimization-based (VO) [13], [14] methods. The CS approaches commonly separate the spatial component and spectral component of the MS image by projecting it into a suitable space, and further substitute the spatial component with the PAN image. The representative CS methods include the band-dependent

The associate editor coordinating the review of this manuscript and approving it for publication was Manuel Rosa-Zurera.

spatial-detail with local parameter estimation (BDSD) [15], the intensity-hue-saturation (IHS) method [16], the principal component analysis (PCA) method [9], [17], and the Gram-Schmidt (GS) spectral sharpening [18]. The sharpened outcomes of CS methods usually show high spatial fidelity paid but suffer from greater spectral distortions [19].

The MRA methods are mainly based on the spatial details injection manner. Specifically, these approaches first extract spatial details from the PAN image and then inject them into the MS image. The products fused by the MRA methods can achieve well-preserved spectral information while possessing obvious spatial distortions even severe artifacts. The most common instances of MRA include smoothing filter-based intensity modulation (SFIM) [20], the additive wavelet luminance proportional (AWLP) [11], decimated wavelet transform (DWT) [21], atrous wavelet transform (ATWT) [22], Laplacian pyramid (LP) [23], the generalized Laplacian pyramid (GLP) [24], and the GLP with full-scale regression (GLP-Reg) [25].

In the last several years, the VO approaches have been widely concerned thanks to their desirable fusion effect on pansharpening. Methods belonging to this class attempt to establish specific optimization functions based on certain conditions [26]. The first VO method treats the PAN image as the linear combination of diverse bands of HRMS image, thus the LRMS image is the blurred version of HRMS image [27]. Afterward, various VO methods are developed to address pansharpening problem, such as Bayesian methods [13], [14], [28], variational approaches [29], [30], [31], compressed-sensing and sparse representation-based techniques [32], [33], [34], [35], [36], [37] and so on. Despite these methods achieving a good balance between spectral information and spatial details by optimizing the loss function, they inevitably introduce more tunable parameters and a higher computational burden.

Recently, deep learning (DL) methods that are mainly based on convolutional neural networks (CNNs) have dominated various image processing tasks, such as object detection [38], [39], [40], image segmentation [41], [42], [43], [44], [45], image super-resolution [46], [47], [48], and presented desirable performance. In the field of pansharpening, Masi et al. [49] firstly introduced a simple CNN architecture inspired by SRCNN [46], dubbed PNN, which consists of only three convolution layers. Yang et al. exploited a deeper network by adopting the residual learning module in Resnet [50] for pansharpening, which achieved significant progress in comparison to traditional pansharpening methods. Afterward, more deeper and complicated CNN architectures are developed for the pansharpening problem. Yuan et al. [51] presented a good technique to inject the high-frequency information into the up-sampled MS image when training the network. DiCNN [52], a detail injection framework, was proposed with two variants (*i.e.,* DiCNN1 and DiCNN2), which conducted the detail injections based on PAN or MS and PAN together, respectively. Deng et al. [19] developed a deep fusion network by borrowing the ideas from the traditional CS and MRA methods, showing favorable fusion results as well as desired generalization capacity. Jin et al. [53] firstly explored the application of a highly anticipated adaptive technique in the pansharpening task and proposed a novel convolution operator by modifying the standard convolution. Due to the effectiveness of the attention mechanism, Liu et al. [54] utilized the attention module and proposed an attention-based network to fuse the PAN and MS image.

Although existing DL-based methods have achieved remarkable progress against their traditional counterpart. Most of them mainly focus on spatial feature extraction while neglecting the significance of spectral characteristics. Instead of only paying attention to the spatial information, we also attempt to explore the implicit relationships in the spectral space within the consideration of spatial features. Specifically, we propose a simple yet effective group shuffle (GS) module to extract the continuous characteristics of various channels while enhancing the long-distance interaction among them. This functional design possesses powerful representation ability and is suitable for the pansharpening problem. Furthermore, we design a simple pansharpening network (dubbed CGSNet) whose backbone is composed of several GS modules, and we further validate its fusion ability on various satellite datasets (i.e. WorldView-3(WV3), GaoFen-2(GF2), QuickBrid(QB) and WorldView-2(WV2) dataset). Extensive experiments demonstrate that our CGSNet achieves competitive fusion results compared with state-of-the-art pansharpening methods in both reduced and full-resolution assessments.

The main contribution of this work can be summarised as follows:

1) Existing deep learning-based pansharpening methods mainly focus on enhancing the spatial representation ability of the network, while paying little attention to modeling spectral dependencies in spite of its significance for remote sensing data. We first attempt to explore the implicit relationships in the spectral space within the consideration of spatial features.
2) We propose a simple yet effective channel group shuffling module which consists of two basic operations: channel grouping and feature shuffle fusion. The first is utilized to enhance the spectral diversity of the modality features. The latter is responsible for integrating the cross-group feature maps with rich spatial-spectral information, thus obtaining more informative features.
3) Extensive experiments show that our CGSNet outperforms other state-of-the-art (SOTA) counterparts over different satellite datasets and the ablation studies further prove the flexibility and effectiveness of the proposed module.

We organize the remaining parts of this paper as follows: Section II introduces the related works and channel shuffling mechanisms. Section III formulates the proposed

methodology. Section IV presents the experimental details, including the datasets simulation, evaluation metrics, experimental settings, comparison experiments, and ablation studies. Section V concludes the proposed method.

II. RELATED WORKS

In this section, we first introduce some representative DL-based pansharpening methods, and then we present the previous channel shuffling mechanism.

A. DL-BASED PANSHARPENING METHODS

Over the past years, DL-based pansharpening methods have shown remarkable progress in comparison to traditional algorithms, which mainly benefit from the excellent nonlinear fitting and feature representation abilities of CNNs. Masi et al. [49] first introduced a CNN network that consists of three convolution layers for pansharpening. Yang et al. [55] developed a deeper CNN architecture, dubbed PanNet, for gathering spatial information. Afterward, deeper and more complicated network architectures have been proposed for pansharpening. To capture spatial features with different scales, Yuan et al. [51] designed a multi-scale-and-depth convolutional neural network (MSDCNN) using the multi-scale convolution block as a basic component. Deng et al. [19] designed an interpretable network for pansharpening by combining the ideas of traditional CS and MRA methods, *i.e.,* FusionNet, which presented a favorable fusion effect and desirable generalization capacity. Jin et al. [53] argued that standard convolution with static kernels suffers from limited representation ability, and proposed an adaptive convolution technique (*i.e.,* LAGConv) whose kernels are generated according to the input patch. Inspired by LAGConv, ADKNet [56] constructed two adaptive kernel generation branches for gathering spatial and spectral information, respectively. Zhou et al. [57] developed a task-specific transformer architecture to capture the long-range spatial features and then introduced the invertible neural module to effectively fuse the obtained features. All of these models mainly focus on learning the complementary information between the MS image PAN image, while rarely considering the implicit relationships among spectral bands. Unlike natural images, however, remote sensing data commonly contain multi-bands that embrace complicated spectral dependencies. Therefore, we attempt to explore the implicit relationships among various spectral bands within the consideration of spatial information to enhance the fusion ability of the network.

B. CHANNEL SHUFFLING OPERATION

Channel shuffling operation was originally developed for lightweight network design. Zhang et al. [58] first investigated the usage of channel shuffle operation in tiny model design and proposed a computation-efficient CNN architecture, called ShuffleNet, for mobile devices. Afterward, channel shuffling operation is increasingly adopted to enhance the cross-channel information interaction. In [59], the channel shuffle operation is implemented on the feature maps with different levels to promote cross-channel information communication among the pyramid feature maps. Huang et al. [60] proposed a novel Shuffle Transformer based on the spatial shuffle to achieve connections among windows. Sun et al. [61] proposed a simple yet effective lightweight image super-resolution network, dubbed ShuffleMixer, which adopts the channel splitting and shuffling operation to achieve the feature interaction. Pang et al. [62] proposed a new Transformer-MLP paradigm, called 3D Shuffle-Mixer, for medical dense prediction. Despite channel shuffling operation being widely used in various vision tasks, there are few works that focus on exploring the implicit relationships in the spectral space that is common and significant for remote sensing images with multi-bands.

III. METHODS

In this section, we first introduce the math notation and then elaborate on the design of our CGSNet with the unique channel group module and feature shuffle fusion module. Finally, we introduce the loss function used to train our CGSNet.

A. MATH NOTATION

Satellites often capture two types of images simultaneously, which are single-band high-resolution panchromatic (PAN) image $\boldsymbol{P} \in \mathbb{R}^{H\times W\times 1}$ and multi-band low-resolution image $\boldsymbol{M} \in \mathbb{R}^{h\times w\times C}$. We need to fuse these two images together and obtain a multi-band super-resolution (SR) image $\boldsymbol{S} \in \mathbb{R}^{H\times W\times C}$. We denote the fused ground truth image as $\boldsymbol{G} \in \mathbb{R}^{h\times w\times C}$ if it exists. The deep neural network for fusion is denoted as $f_\theta(\cdot)$ with parameters θ. Usually, the fused process can be defined as follows:

$$\boldsymbol{S} = f_\theta(\boldsymbol{M}, \boldsymbol{P}). \tag{1}$$

When there exists the ground-truth fused image, a supervised loss can be formed to enable the gradient descent. After many training epochs, the network is trained and converged.

B. OVERALL ARCHITECTURE

As shown in Fig. 1, we propose a simple and effective network, Channel Group Shuffle Network (CGSNet), which contains two main modules, channel group module (CG) and feature shuffle fusion (FSF) module. The CG module groups the spectral information in successive and interval manners along the feature channel dimension, and the FSF module is responsible for fusing grouped spectral information back into a feature map. The corporation of these two modules can help the network gather the spectral information explicitly (which is important for multi-spectral images) and the spatial information propagation is guaranteed by the simple fusion head.

First, the PAN image $\boldsymbol{P}$ and upsampled LRMS $\hat{\boldsymbol{M}}$ are concatenated along the channel dimension and fed into the CG module. After performing grouping, we get two different grouped feature maps. Then, the two feature maps

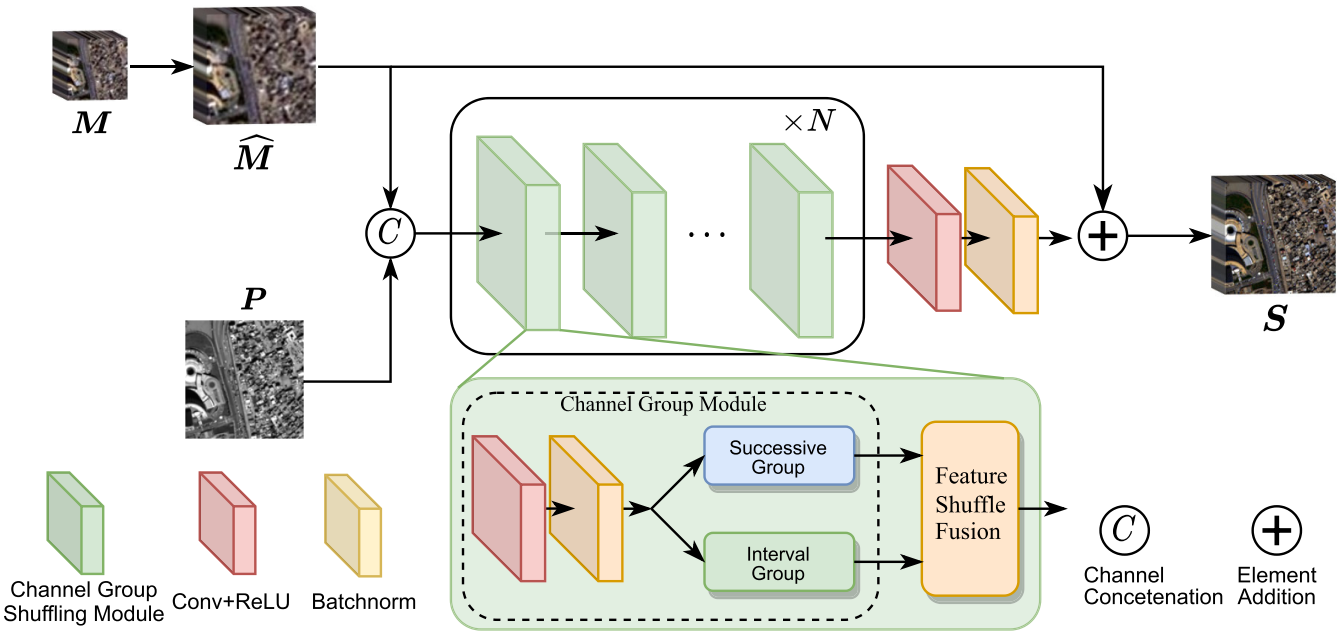

FIGURE 1. Our CGSNet contains two modules, which are the channel group shuffle module and the channel feature fusion module, respectively. The fusion head close to the output is simple, which is a combination of convolution, ReLU, and batchnorm. These two modules are stacked *N* layers for full information exchange.

are sent into the FSF module to fusion and propagate spectral information back to a single feature map, which completes long-term dependency construction. Finally, the residual patch from the input to the output is added for learning high-frequency components which is proven to be useful. The output of the network $\boldsymbol{S} = f_\theta(\hat{\boldsymbol{M}}, \boldsymbol{P}) + \hat{\boldsymbol{M}}$ is used for computing loss function with the ground-truth (GT) images.

C. CHANNEL GROUP MODULE

The Channel Group (CG) module is the main module of our CGSNet. It explicitly groups the feature maps along the channel dimension. Traditional 1×1 convolution can only respond to the feature just in its kernel, regardless of the features outside. Sliding along the channel dimension only expands its receptive field but does not help with responding two points far away. Aware of this, the CG module groups the features successively and intervally, which can be denoted as follows,

$$\mathbf{GP_{su}} = \{\mathbf{GP}^i_{\mathbf{su}}\} = \{[F_0, F_1, \ldots, F_{g-1}], \ldots, [F_{(i-1)g}, F_{(i-1)g+1}, \ldots, F_{ig-1}]\}, 1 \le i \le I, \quad (2)$$

$$\mathbf{GP_{in}} = \{\mathbf{GP}^i_{\mathbf{in}}\} = \left\{\left[F_0, F_g, \ldots, F_{\lfloor \frac{d}{I} \rfloor}\right], \ldots, \left[F_{d-g}, F_{d-g+1}, \ldots, F_{d-1}\right]\right\}, 1 \le i \le I, \quad (3)$$

where i is the group index. I denotes the number of groups of successive groups $\mathbf{GP_{su}}$ and interval groups $\mathbf{GP_{in}}$. F is the input feature maps with d channel dimensions. Successive group operation maintains a similar feature gathering manner as 1×1 convolution, and interval grouping operation gathers two channels together by an interval g.

These two group operations can be implemented really neatly and easily with the pseudocode shown in Alg.1.

Algorithm 1 Python-Like Pseudocode for Successive and Interval Grouping Operation

```
# F:input feature maps
# d:the number of dimension
# ngroups:the number of groups
def CG(F, d, ngroups)
    # Successive group operation
    GP_su ← Split(F, d // ngroups,
     dim=1)
    # Interval group operation
    GP_in ← EmptyList()
    for i in range(ngroups) do
     | GP_in ← append(F[:, i::ngroups])
    end
    return GP_su, GP_in
```

By grouping different channels of the feature maps into successive and interval groups, the consistency of channels and long-range dependency are guaranteed, respectively.

D. FEATURE SHUFFLE FUSION MODULE

CG module only benefits feature gathering but without shuffling the channel information, so it can not help propagate the information, we then introduce the feature shuffle fusion (FSF) module as shown in Fig.2. Given two groups $\mathbf{GP_{su}}, \mathbf{GP_{in}}$ gathered above, we reshuffle the two groups channel-wisely and form the new I groups $\widehat{\mathbf{GP}} = \{GP_i\}$,

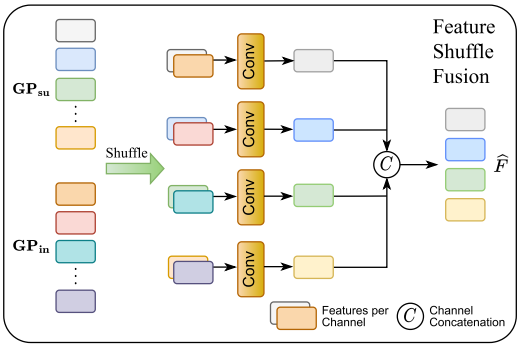

FIGURE 2. Feature shuffle fusion module. First, the successive groups and interval groups are shuffled channel-wisely into another I groups. Then, several convolution layers are performed on these shuffled groups to propagate information. Finally, the fused feature maps $\hat{\mathbf{F}}$ are produced by concatenating reshuffled groups along channel dimension.

$0 \leq i \leq I$, which can be denoted as follows,

$$\widehat{\mathbf{GP}}^i = \mathbf{GP}^{\mathbf{i}}_{\mathbf{su}} \copyright \mathbf{GP}^i_{\mathbf{in}}, \tag{4}$$

$$\hat{\mathbf{F}}_i = \text{Conv}(\mathbf{GP}^i), \tag{5}$$

where $\copyright$ means channel concatenation, Conv represents a resblock [50] and $\hat{\mathbf{F}}_i$ denote the i^{-th} channel of the fused feature map $\hat{\mathbf{F}}$. First, successive groups and interval groups are extracted channel-wisely and shuffled into I new groups $\widehat{\mathbf{GP}}^i$. Then, to propagate the information from the new groups, a weight-sharing resblock is adopted. Finally, we concatenate the groups along the channel dimension to form a fused feature map $\hat{\mathbf{F}}$.

E. LOSS FUNCTION

Most previous works directly employ an L1 reconstruction loss, which is spectrally adequate. For better fusion quality and fidelity, we consider a combination of L1 loss and SSIM loss for PSNR and SSIM (metric) orientation optimization, respectively. The total loss function can be represented as follows:

$$\mathcal{L}_1 = \frac{1}{HWC} \sum_{i,j,c} \|\boldsymbol{S}(i,j,c) - \boldsymbol{G}(i,j,c)\|_1, \tag{6}$$

$$\mathcal{L}_{ssim} = 1 - SSIM(\boldsymbol{S}, \boldsymbol{G}), \tag{7}$$

$$\mathcal{L}_{total} = \mathcal{L}_1 + 0.1\mathcal{L}_{ssim}, \tag{8}$$

where H, W, C are the height, width, and the number of channels of the image. The L1 loss optimizes the PSNR and the SSIM loss can better fuse the details. $\mathcal{L}_{total}$ is a weighted loss of these two losses by a factor of 0.1. We ablate the choices of the loss function and the weighted factor in Sec. IV-H2.

IV. EXPERIMENTS

In this section, we will introduce the datasets, experiment settings, the performance of our CGSNet, and comparisons of other state-of-the-art methods.

A. DATASETS

We conduct extensive experiments on three widely used remote sensing datasets, i.e. WorldView-3 (WV3), GaoFen-2 (GF2), and QuickBrid (QB).

- WV3 dataset contains 9714/1080 train/validation pairs, each pairs are composed of PAN, LRMS, and HRMS images.
- GF2 dataset contains 19809/2201 train/validation pairs, each pairs are composed of PAN, LRMS, and HRMS images.
- QB dataset contains 17139/1905 train/validation pairs, each pairs are composed of PAN, LRMS, and HRMS images.

We treat HRMS images as GT images. For testing, there are 20 pairs for examining reduced-resolution fusing capability and another 20 pairs without the GT for testing full-resolution fusing performance.

The training MS has the shape of $16 \times 16 \times c$ ($c = 8$ for WV3 and 4 for GF2 and QB), and the training PAN has the shape of $64 \times 64 \times 1$. The validation MS and PAN maintain the same shape. The reduced-resolution MS has the shape of $64 \times 64 \times c$, while the reduced-resolution PAN has the shape of $256 \times 256 \times 1$. The full-resolution MS has the shape of $128 \times 128 \times c$, while the reduced-resolution PAN has the shape of $512 \times 512 \times 1$.

B. EVALUATION METRIC

To evaluate the performances of our CGSNet and other methods, we used seven widely-used quality metrics including SAM [63], ERGAS [64], Q2n and SCC [65] for reduced-resolution metrics, while Q_λ, Q_s and QNR [66] for full-resolution metrics. The detailed mathematical formula of these metrics are described as follows:

1) SPECTRAL ANGLE MAPPER (SAM)

The SAM metric calculates the angle between the fused image and the HRMS image to evaluate the spectral quality. The ideal value is 0. The SAM metric can be computed as follows:

$$\text{SAM} = \frac{1}{C} \sum_{i=1}^{C} \arccos\left(\frac{X_i \cdot \hat{X}_i}{\|X_i\|_2 \|\hat{X}_i\|^2}\right), \tag{9}$$

where C represents the number of spectral, X_i and $\hat{X}_i$ denote i-th spectral vector of the GT and the fused image. $\|\cdot\|_2$ means the L2 norm.

2) RELATIVE DIMENSIONLESS GLOBAL ERROR IN SYNTHESIS (ERGAS)

ERGAS is a metric for evaluating the overall synthesis effect of multispectral images, which combines spatial and frequency spectral information, enabling a more comprehensive assessment of the quality of synthesized images. The ideal value for ERGAS is 0. ERGAS can be expressed as the

following formula:

$$\mathrm{ERGAS}(X,\hat{X}) = \frac{100}{s}\sqrt{\frac{1}{C}\sum_{i=1}^{C}\frac{\mathrm{MSE}(X_i,\hat{X}_i)}{\mu_{\hat{X}_i^2}}}. \tag{10}$$

Among them, s represents the spatial downsampling ratio, X is the GT image, and $\hat{X}$ is the fused image. $\mathrm{MSE}(X_i,\hat{X}_i)$ represents the mean square error between X_i and $\hat{X}_i$. $\mu_{\hat{X}_i^2}$ represents the root mean square error of $\hat{X}_i$.

3) SPATIAL CORRELATION COEFFICIENT (SCC)

SCC is used to evaluate the similarity of the spatial details of the fused image and the GT image through a high-pass filter and to calculate the correlation coefficient (CC). The high-pass filter is formed as follows,

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{11}$$

The ideal value of SCC is 1, the higher value means the higher spatial similarity between the fused image and the GT image. The CC is calculated as follows,

$$CC = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j}-\mu_X)(\hat{X}_{i,j}-\mu_{\hat{X}})}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j}-\mu_X)^2(\hat{X}_{i,j}-\mu_{\hat{X}})^2}}, \tag{12}$$

where X is the GT image and $\hat{X}$ is the fused image. w and h are the width and height of the image, respectively. μ denotes the mean value of the image.

4) QUALITY INDEX (Q2N)

Metric Q2n combines three factors to calculate image distortion, which are correlation loss, brightness distortion, and contrast distortion. The Q function is defined as follows,

$$Q = \frac{|\sigma_{Z_1,Z_2}|}{\sigma_{Z_1}\cdot\sigma_{Z_2}}\cdot\frac{2\sigma_{Z_1}\cdot\sigma_{Z_2}}{\sigma_{Z_1}^2+\sigma_{Z_2}^2}\cdot\frac{2|\bar{Z}_1|\cdot|\bar{Z}_2|}{|\bar{Z}_1|^2\cdot|\bar{Z}_2|^2}, \tag{13}$$

where Z_1 and Z_2 represent the b-th band of the fused image and the GT image, respectively. When Q2n is 1, this represents the best fidelity. Q2n metric is defined following [67]. When the spectral number is 8 (e.g. WorldView-3 dataset), Q2n is Q8. When the spectral number is 4 (e.g. GaoFen2 dataset), Q2n is Q4.

5) SPECTRAL DISTORTION INDEX (D_λ)

The spectral distortion index D_λ measures the degree of image distortion in the frequency domain, mainly considering the color information of the image. The calculation formula of D_λ is as follows,

$$D_\lambda = \sqrt[q]{\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N}|d_{i,j}(\boldsymbol{M},\hat{\boldsymbol{M}})|^q}, \tag{14}$$

where N is the number of pixels and $d_{i,j} = Q(\boldsymbol{M}_i,\boldsymbol{M}_j) - Q(\hat{\boldsymbol{M}}_i,\hat{\boldsymbol{M}}_j)$.

6) SPATIAL DISTORTION INDEX (Q_S)

The spatial distortion index D_s measures the degree of image distortion in space, considering the edge texture information of the image, the calculation formula is as follows,

$$D_s = \sqrt[q]{\frac{1}{N}\sum_{i=1}^{N}|Q(\hat{\boldsymbol{M}}_i,\boldsymbol{P}) - Q(\boldsymbol{M}_i,\boldsymbol{P}_{LS})|^q}, \tag{15}$$

where, $\boldsymbol{P},\boldsymbol{M}$ represent the PAN image and LRMS image respectively, and $\boldsymbol{P}_{LS}$ is the low resolution PAN image downsampled by r times. q is usually set to 1.

7) QUALITY W/O REFERENCE METRIC (QNR)

The QNR index is defined as,

$$\mathrm{QNR} = (1-D_\lambda)^\alpha(1-D_s)^\beta, \tag{16}$$

where usually $\alpha = \beta = 1$. The QNR metric can reflect the spectral and spatial distortion of the fused image.

C. BENCKMARK

For comparisons, we choose three widely used traditional methods and SOTA DL-based methods listed as follows:

- Traditional methods: BDSD-PC [68], MTF-GLP-FS [25] and BT-H [69]
- DL-based methods: PNN [49], PanNet [55], DiCNN [52], MSDCNN [51], FusionNet [19] and CTINN [57].

D. EXPERIMENT SETTING

We conduct our experiments with the Pytorch deep learning package with one 3090 Nvidia GPU. We use the AdamW optimizer and set the initial learning rate to 0.001 then halve it in multistep with steps 300 and 800. We train the CGSNet for 1000 epochs with a batch size of 64. We set $N = 1$ and found one grouping and shuffling are adequate and g is empirically set to 16. We will re-examine this setting in the ablation studies.

E. REDUCED ASSESSMENTS

In the WV3 reduced-resolution test set, compared with previous SOTA methods, it can be found that our CGSNet reached a new SOTA performance on SAM, ERGAS, and Q4 metrics and a competitive performance on SCC metric in Tab. 1. The traditional methods are less competitive than all DL-based methods. When comparing with the previous best method CTINN [57], our CGSNet owns a 3.20/2.37/0.91/0.9826 better SAM, ERGAS, Q4, and SCC metrics. The fused HRMS of CGSNet have clear boundaries and less spatial distortion, which can be seen in the boundaries of buildings in Fig. 3. The error map with the GT clearly shows that CGSNet is closer to the GT as it has the darkest color.

For GF2 reduced-resolution fusion, our CGSNet still owns the best performances on all metrics. The SAM, ERGAS, and Q4 metrics of CGSNet are improved by ≈7.3%/5.2%/0.87%. For a clear visual comparison, the fused HRMS and the

TABLE 1. Average quantitative metrics on 20 examples for the WV3 dataset. Some conventional methods (the first three rows) and CNN methods are compared. (Bold: best; Underline: second best).

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q4(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
BDSD-PC	5.4675±1.7185	4.6549±1.4667	0.8117±0.1063	0.9049±0.0419	0.0281±0.0171	0.0730±0.0356	0.9009±0.0474
MTF-GLP-FS	5.3233±1.6548	4.6452±1.4441	0.8177±0.1014	0.8984±0.0466	0.0354±0.0211	0.0630±0.0284	0.9043±0.0454
BT-H	4.8985±1.3028	4.5150±1.3315	0.8182±0.1019	0.9240±0.0243	0.0430±0.0232	0.0810±0.0374	0.8803±0.0540
PNN	3.6798±0.7625	2.6819±0.6475	0.8929±0.0923	0.9761±0.0075	0.0423±0.0080	0.0528±0.0216	0.9071±0.0212
PanNet	3.6156±0.7665	2.6660±0.6887	0.8906±0.0934	0.9757±0.0088	0.0285±0.0084	0.0473±0.0229	0.9255±0.0271
DiCNN	3.5929±0.7623	2.6733±0.6627	0.9004±0.0871	0.9763±0.0072	0.0362±0.0111	0.0462±0.0175	0.9195±0.0258
MSDCNN	3.7773±0.8032	2.7608±0.6884	0.8900±0.0900	0.9741±0.0076	**0.0230±0.0091**	0.0467±0.0199	0.9316±0.0271
FusionNet	3.3252±0.6978	2.4666±0.6446	0.9044±0.0904	0.9807±0.0069	0.0339±0.0092	0.0621±0.0241	0.9061±0.0197
CTINN	3.2523±0.6436	2.3936±0.5194	0.9056±0.0840	**0.9826±0.0046**	0.0550±0.0288	0.0679±0.0312	0.8815±0.0488
CGSNet	**3.2024±0.6848**	**2.3772±0.6167**	**0.9136±0.0860**	0.9824±0.0063	0.0242±0.0082	**0.0437±0.0170**	**0.9332±0.0234**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

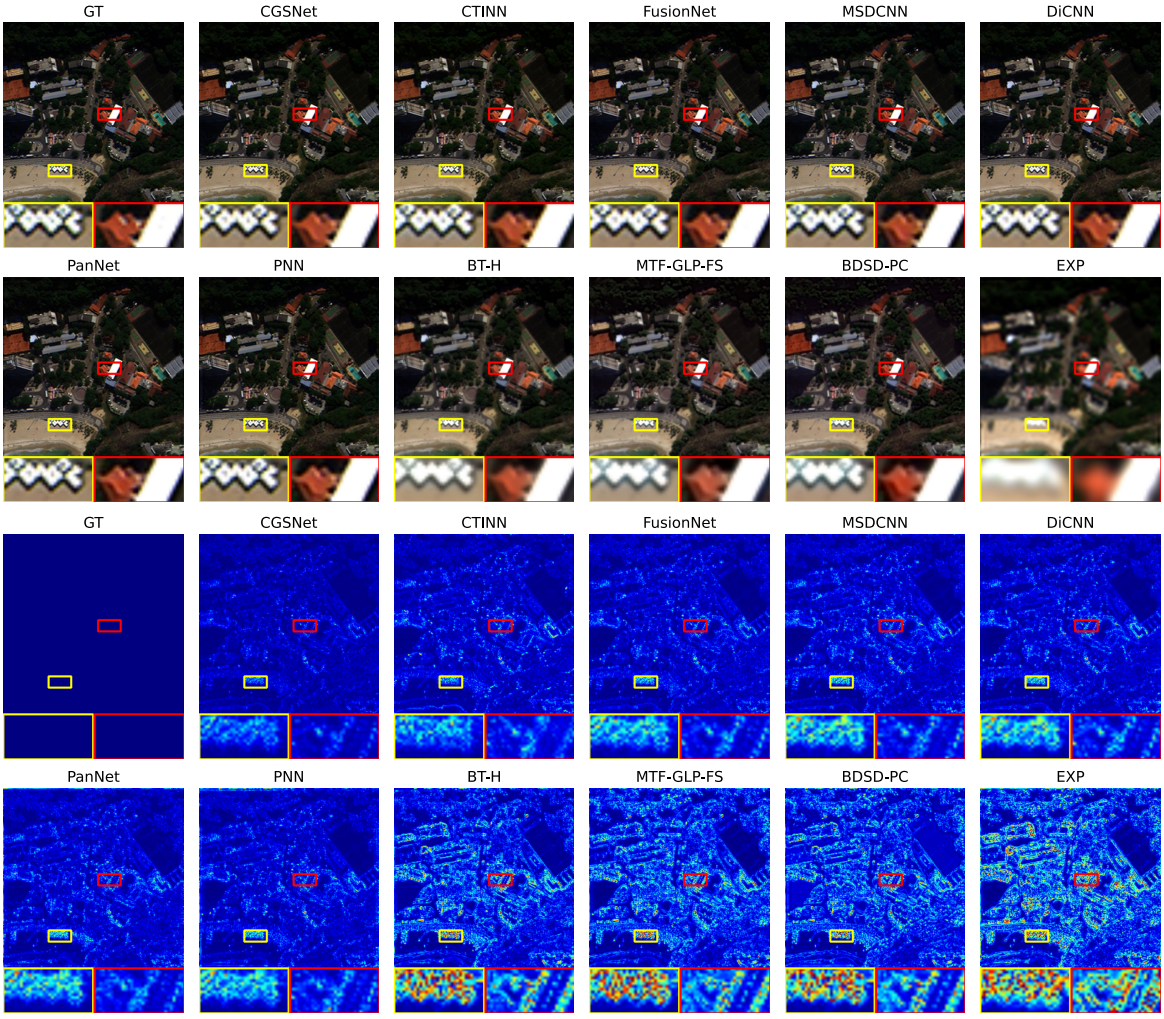

FIGURE 3. Visual comparisons of the fused HRMS from traditional methods and previous SOTA DL-based methods on WV3 reduced-resolution test set. The upper panel is the fused HRMS and GT. The lower panel is the error maps compared with the GT.

error maps with the GT are illustrated in Fig. 4. It is clear that CGSNet can fuse a clearer HRMS with less spatial and spectral error.

When it comes to QB reduced-resolution test set, CGSNet still performs well. The quality metrics of CGSNet are better than previous SOTA methods. The previous best method CITNN lags behind 0.05/0.02/0.012 on SAM/ERGAS/Q4 metrics. Similarly, we plot the visualizations of fused images. Our CGSNet fuses distortion-free HRMS images with better visual qualities as shown in Fig. 5.

For the reduced-resolution assessments, SAM and Q2n metrics significantly represent the spectral distortions of the

TABLE 2. Average quantitative metrics on 20 examples for the GF2 dataset. Some conventional methods (the first three rows) and CNN methods are compared. (Bold: best; Underline: second best).

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q4(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
BDSD-PC	1.7110±0.3210	1.7025±0.4056	0.9932±0.0308	0.9448±0.0166	0.0759±0.0301	0.1548±0.0280	0.7812±0.0409
MTF-GLP-FS	1.6757±0.3457	1.6023±0.3545	0.8914±0.0256	0.9390±0.0197	0.0759±0.0301	0.1548±0.0280	0.7812±0.0409
BT-H	1.6810±0.3168	1.5524±0.3642	0.9089±0.0292	0.9508±0.0150	0.0602±0.0252	0.1313±0.0193	0.8165±0.0305
PNN	1.0477±0.2264	1.0572±0.2355	0.9604±0.0100	0.9772±0.0054	0.0367±0.0291	0.0943±0.0224	0.8726±0.0373
PanNet	0.9967±0.2119	0.9192±0.1906	0.9671±0.0099	0.9829±0.0035	**0.0206±0.0112**	0.0799±0.0178	0.9011±0.0203
DiCNN	1.0525±0.2310	1.0812±0.2510	0.9594±0.0101	0.9771±0.0058	0.0413±0.0128	0.0992±0.0131	0.8636±0.0165
MSDCNN	1.0472±0.2210	1.0413±0.2309	0.9612±0.0108	0.9782±0.0050	0.0369±0.0131	**0.0730±0.0093**	0.8927±0.0128
FusionNet	0.9735±0.2117	0.9878±0.2222	0.9641±0.0093	0.9806±0.0049	0.0400±0.0126	0.1013±0.0134	0.8628±0.0184
CTINN	0.8981±0.1396	0.7995±0.1292	0.9689±0.0137	0.9803±0.0015	0.0586±0.0260	0.1096±0.0149	0.8381±0.0237
CGSNet	**0.8323±0.1691**	**0.7580±0.1467**	**0.9773±0.0080**	**0.9887±0.0221**	0.0290±0.0120	0.0800±0.0111	**0.8933±0.0157**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

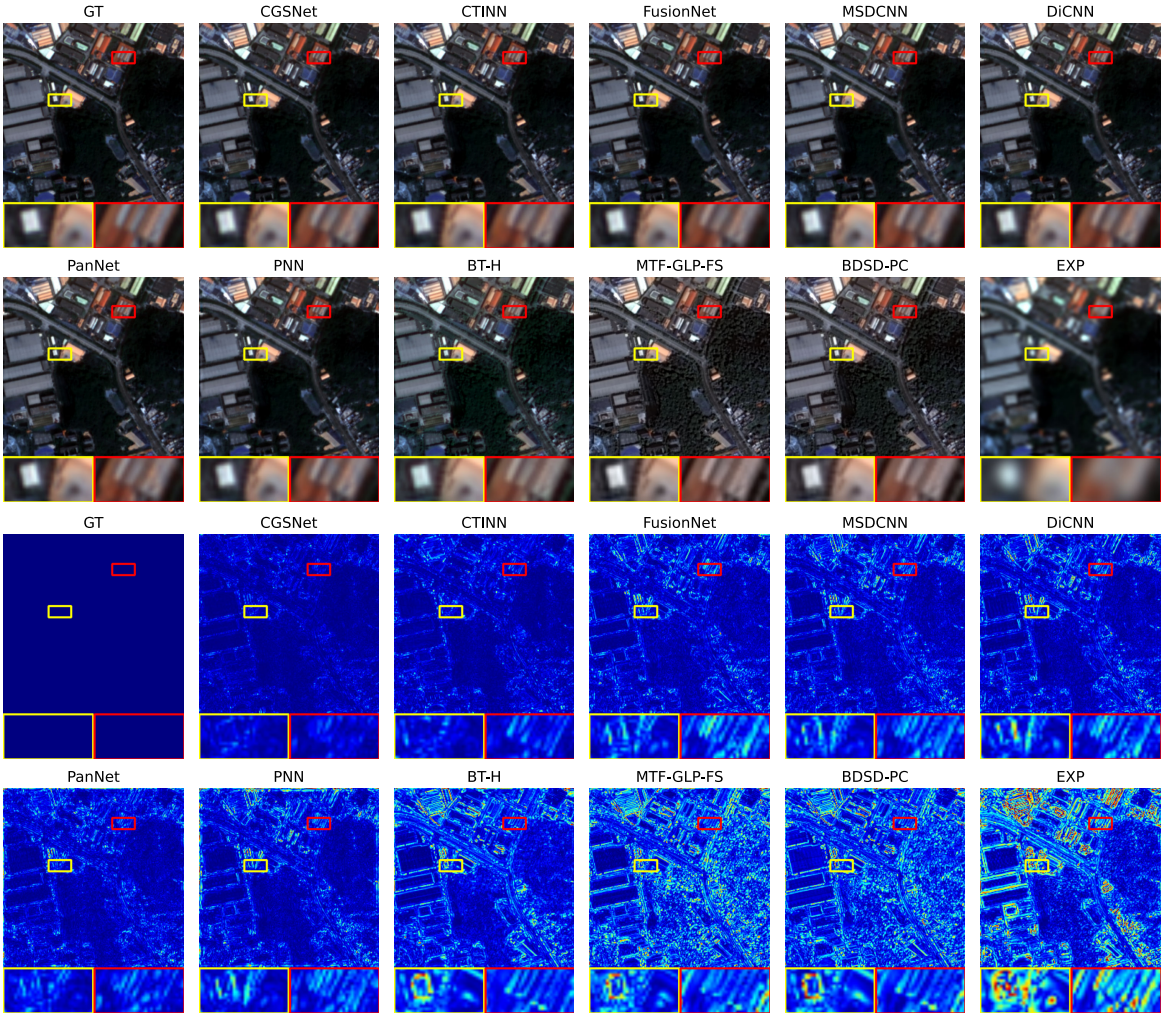

FIGURE 4. Visual comparisons of the fused HRMS from traditional methods and previous SOTA DL-based methods on GF2 reduced-resolution test set. The upper panel is the fused HRMS and GT. The lower panel is the error maps compared with the GT.

fused images. As presented in the left panels of Tab. 1, Tab. 2, and Tab. 3, our CGSNet obtains the optimal SAM and Q2n values in comparison to other methods, showing its superiority in modeling the spectral dependencies of the input modality features.

F. FULL ASSESSMENTS

To verify the generalization quality of the proposed method, the non-reference performance assessments on the full-resolution test set can be used to represent. We separately test the quality assessments on WV3,

TABLE 3. Average quantitative metrics on 20 examples for the QB dataset. Some conventional methods (the first three rows) and CNN methods are compared. (Bold: best; Underline: second best).

method	Reduced				Full		
	SAM(± std)	ERGAS(± std)	Q4(± std)	SCC(± std)	D_λ(± std)	D_s(± std)	QNR(± std)
BDSD-PC	8.2620±2.0497	7.5420±0.8138	0.8323±0.1013	0.9030±0.0181	0.0345±0.0172	0.1636±0.0483	0.8078±0.0497
MTF-GLP-FS	8.1131±1.9553	7.5102±0.7926	0.8296±0.0905	0.8998±0.0196	0.0570±0.0137	0.1500±0.0238	0.8017±0.0295
BT-H	7.1943±1.5523	7.4008±0.8378	0.8326±0.0880	0.9156±0.0152	0.0526±0.0141	0.1648±0.0167	0.7912±0.0177
PNN	5.2054±0.9625	4.4722±0.3734	0.9180±0.0938	0.9711±0.0123	0.0569±0.0112	0.0624±0.0239	0.8844±0.0304
PanNet	5.7909±1.1839	5.8629±0.8883	0.8850±0.0917	0.9485±0.0170	0.0410±0.0108	0.1137±0.0323	0.8502±0.0390
DiCNN	5.3795±1.0266	5.1354±0.4876	0.9042±0.0942	0.9621±0.0133	0.0920±0.0143	0.1067±0.0210	0.8114±0.0310
MSDCNN	5.1471±0.9342	4.3828±0.3400	0.9176±0.0987	0.9722±0.0124	**0.0320±0.0237**	0.0667±0.0282	0.9041±0.0466
FusionNet	4.9226±0.9077	4.1594±0.3212	0.9252±0.0902	0.9755±0.0104	0.0586±0.0189	0.0522±0.0088	0.8922±0.0219
CTINN	4.6583±0.7755	3.8969±0.2888	0.9320±0.0072	0.9829±0.0072	0.1738±0.0332	0.0731±0.0237	0.7663±0.0432
CGSNet	**4.6124±0.8413**	**3.8841±0.3090**	**0.9335±0.0843**	**0.9854±0.0076**	0.0439±0.0108	**0.0428±0.0144**	**0.9154±0.0227**
Ideal value	**0**	**0**	**1**	**1**	**0**	**0**	**1**

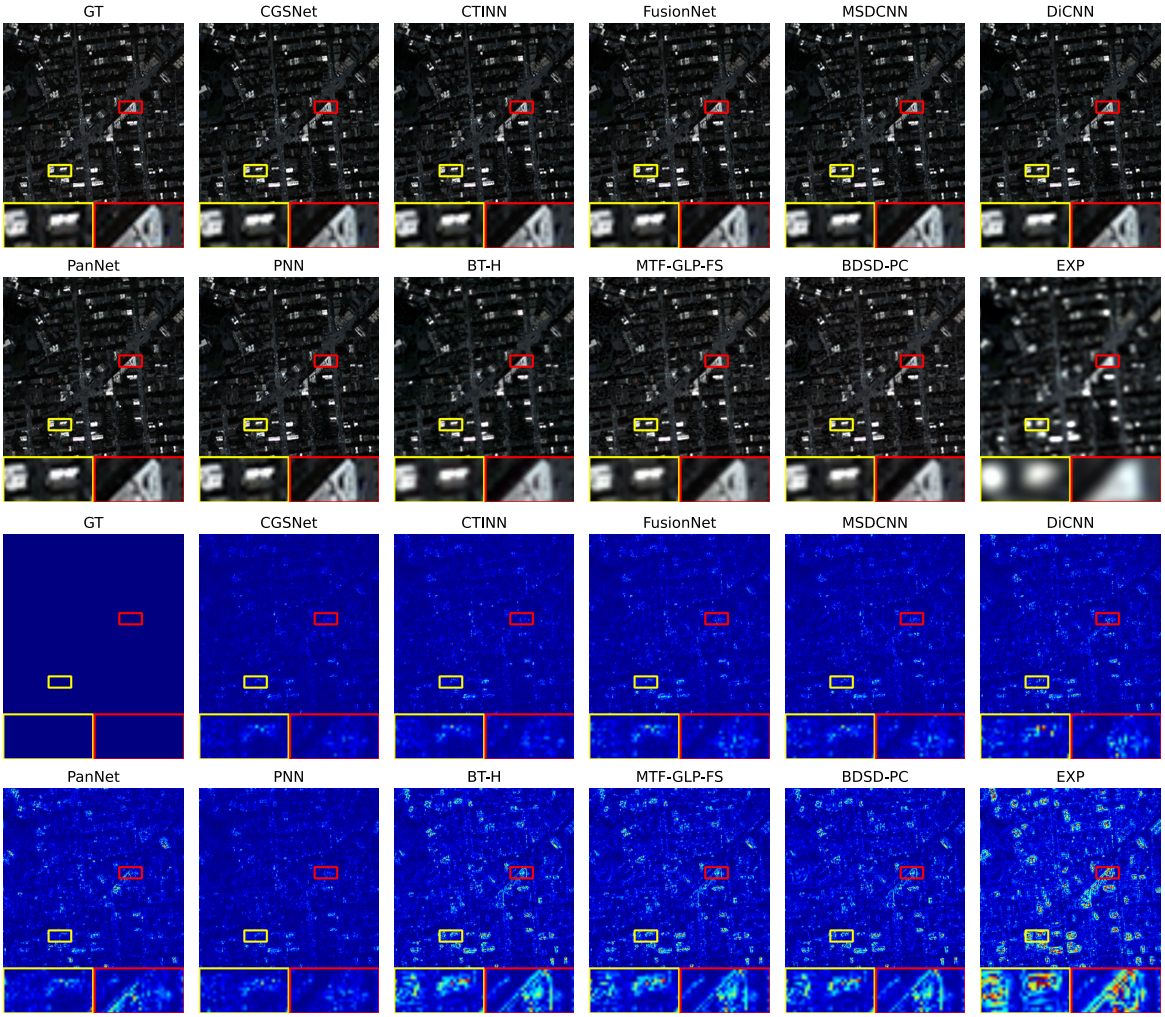

FIGURE 5. Visual comparisons of the fused HRMS from traditional methods and previous SOTA DL-based methods on QB reduced-resolution test set. The upper panel is the fused HRMS and GT. The lower panel is the error maps compared with the GT.

GF2, and QB full-resolution test sets after obtaining models trained on the reduced-resolution dataset. Since the model has a train-test resolution gap, the better-fused performances indicate a better generalization ability.

We test the full-resolution performance on the WV3 test set. As the D_λ and D_s contribute equally to the QNR metric [66], the QNR metric can be considered as the dominant quality metric. As can be seen in Tab. 1, our CGSNet can outperform our previous SOTA methods,

including traditional and DL-based methods. The D_λ metric only lags behind 0.0012 which is relatively small. The D_s metric of the proposed method still outperforms all the previous SOTA method.

As for GF2 full-resolution test set, the quality metrics are reported in Tab. 2. We can see that our CGSNet performs the best on the QNR metric (i.e. 0.8933 compared with the previous SOTA 0.8927). The other two non-reference metrics D_λ and D_s are still competitive.

For QB dataset, the SOTA performances are still obtained by our CGSNet, as the QNR metric reaches 0.9154 while comparing the previous SOTA 0.9041.

For the full-resolution evaluation, the QNR indicator indicates the overall fusion performance including both spatial and spectral qualities of the fused products. It is clearly shown in the right panels of Tab. 1, Tab. 2, and Tab. 3, our model yields the best results on all three datasets, demonstrating its desirable spatial and spectral preservation.

G. ABLATION STUDY

In this section, we conduct an ablation study on the effectiveness of CG, FSF modules, group interval g, loss functions, and the number of stacked layers N to verify their effectiveness.

H. ABLATION ON CG AND FSF MODULE

The effectiveness of the proposed module can be verified by an ablation study. We design another two type of variants of CGSNet:

1) **variant 1**, remove the CG module;
2) **variant 2**, replace the shuffling in the FSF module by simply concatenating.

Then, we train the two variants on the WV3 dataset and test their performances which are reported in Tab. 4. The Variant 1 net performs absolutely worse than the Variant 2 net and the default net because the long-range dependency is not obtained. Variant 2 is still worse since it does not fully use the gathered channel information. This ablation denotes our designed modules can take their effect correctly.

TABLE 4. Ablation on the proposed modules. The gray background means the default setting.

networks	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(±std)
variant 1	3.5992±0.7921	2.5209±0.6920	0.8872±0.0899	0.9702±0.0070
variant 2	3.2976±0.6723	2.3867±0.6390	0.9042±0.0872	0.9782±0.0067
default	**3.2024±0.6848**	**2.3772±0.6167**	**0.9136±0.0860**	**0.9824±0.0063**

1) ABLATION ON GROUP INTERVAL G

We ablate the number of groups g to verify the empirical choice in our main setting is effective. We choose g to be 8, 16, or 32 and conduct the fusing experiment on the WV3 dataset, respectively (see in Tab. 5).

It is easy to find that our empirical setting ($g = 16$) outputs the best fusing results. It is supposed that small g can not obtain enough long-range dependency, and too large g harms the local information.

TABLE 5. Ablation on the choices of interval g. The gray background means the default setting.

g	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(±std)
8	3.3211±0.7201	2.4289±0.6231	0.8992±0.0881	0.9811±0.0061
16	**3.2024±0.6848**	**2.3772±0.6167**	**0.9136±0.0860**	**0.9824±0.0063**
32	3.2291±0.7720	2.4098±0.6109	0.9098±0.0820	0.9802±0.0066

2) ABLATION ON LOSS FUNCTION

To verify the effectiveness of the chosen loss function Eq. 6, we design the ablation study on several loss functions, whose performances on the WV3 dataset are presented in Tab. 6. The default setting performs the best, besides, only with $\mathcal{L}_1$ loss performs the worst due to lack of details supervised loss. Large SSIM loss setting performs less satisfactorily because its gradients domain the overall gradients which hinders the optimization process.

TABLE 6. Ablation on choices of the loss functions. The gray background means the default setting.

loss function	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(±std)
$\mathcal{L}_1$	3.2929±0.7022	2.4002±0.6120	0.9101±0.0799	0.9801±0.0070
$\mathcal{L}_1 + 0.1\mathcal{L}_{\text{ssim}}$	**3.2024±0.6848**	**2.3772±0.6167**	**0.9136±0.0860**	**0.9824±0.0063**
$\mathcal{L}_1 + \mathcal{L}_{\text{ssim}}$	3.3192±0.6929	2.4227±0.6089	0.8980±0.0816	0.9787±0.0062

3) ABLATION ON THE NUMBER OF STACKED MODULES

We propose CG and FSF modules for modeling channel long-range dependency and propagating information. Then we stack the two modules together to form our CGSNet. We set the number of stacked layers to 1 ($N = 1$). In this section, we ablate the number of layers N to 1 (default setting), 3, and 5. As shown in Tab. 7, CGSNet with $N = 5$ obtains the best performances, but the improvement can be neglected when compared to its increases of parameters. In practice, we set $N = 1$ to take the trade-off.

TABLE 7. Ablation on the choices of the number of stacked layers N. The gray background means the default setting.

N	SAM(± std)	ERGAS(± std)	Q8(± std)	SCC(±std)
1	3.2024±0.6848	2.3772±0.6167	0.9136±0.0860	0.9824±0.0063
3	3.1962±0.6719	2.3691±0.6023	0.9140±0.0822	0.9830±0.0064
5	**3.1942±0.6820**	**2.3598±0.6562**	**0.9146±0.0921**	**0.9835±0.0064**

V. CONCLUSION

In this work, we propose a novel channel group shuffling network, termed as CGSNet. Targeted at modeling spectral relationships while preserving spatial information, two core operations are devised to construct the image fusion network: channel grouping operation and cross-group feature fusion operation. Specifically, the former enhances the diversity of spectral information and cross-channel information communications, meanwhile ensuring the spectral order of the input feature, benefiting the pansharpening task. The latter integrates the cross-group feature maps with rich spatial-spectral information. Extensive experiments show that our

CGSNet is capable of outperforming existing state-of-the-art over various satellite datasets.

REFERENCES

[1] A. Sekrecka, M. Kedzierski, and D. Wierzbicki, "Pre-processing of panchromatic images to improve object detection in pansharpened images," *Sensors*, vol. 19, no. 23, p. 5146, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/23/5146

[2] A. A. Alesheikh, A. Ghorbanali, and N. Nouri, "Coastline change detection using remote sensing," *Int. J. Environ. Sci. Technol.*, vol. 4, no. 1, pp. 61–66, Dec. 2007.

[3] V. Walter, "Object-based classification of remote sensing data for change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 58, nos. 3–4, pp. 225–238, Jan. 2004.

[4] T. Blaschke, S. Lang, E. Lorup, J. Strobl, and P. Zeil, "Object-oriented image processing in an integrated GIS/remote sensing environment and perspectives for environmental applications," *Environ. Inf. Planning, Politics Public*, vol. 2, pp. 555–570, Oct. 2000.

[5] G. Tmušić, S. Manfreda, H. Aasen, M. R. James, G. Gonçalves, E. Ben-Dor, A. Brook, M. Polinova, J. J. Arranz, J. Mészáros, R. Zhuang, K. Johansen, Y. Malbeteau, I. P. de Lima, C. Davids, S. Herban, and M. F. McCabe, "Current practices in UAS-based environmental monitoring," *Remote Sens.*, vol. 12, no. 6, p. 1001, Mar. 2020.

[6] K. E. Joyce, S. E. Belliss, S. V. Samsonov, S. J. McNeill, and P. J. Glassey, "A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters," *Prog. Phys. Geography, Earth Environ.*, vol. 33, no. 2, pp. 183–207, Apr. 2009.

[7] C. Corbane, S. Lang, K. Pipkins, S. Alleaume, M. Deshayes, V. E. G. Millán, T. Strasser, J. Vanden Borre, S. Toon, and F. Michael, "Remote sensing for mapping natural habitats and their conservation status—New opportunities and challenges," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 37, pp. 7–16, May 2015.

[8] P. Boccardo and F. G. Tonolo, "Remote sensing role in emergency mapping for disaster response," *Engineering Geology for Society and Territory*, vol. 5. Berlin, Germany: Springer, 2015, pp. 17–24.

[9] X. P. S. Chavez and A. Y. Kwarteng, "Extracting spectral contrast in Landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, Jan. 1989.

[10] J. Qu, Y. Li, and W. Dong, "Hyperspectral pansharpening with guided filter," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2152–2156, Nov. 2017.

[11] 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.

[12] 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, no. 5, pp. 591–596, May 2006.

[13] X. He, L. Condat, J. M. Bioucas-Dias, J. Chanussot, and J. Xia, "A new pansharpening method based on spatial and spectral sparsity priors," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4160–4174, Sep. 2014.

[14] T. Wang, F. Fang, F. Li, and G. Zhang, "High-quality Bayesian pansharpening," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[15] 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.

[16] W. J. Carper, T. M. Lillesand, and R. W. Kiefer, "The use of intensity-hue-saturation transformations for merging SPOT panchromatic and multispectral image data," *Photogram. Eng. Remote Sens.*, vol. 56, no. 4, pp. 459–467, 1990.

[17] 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.

[18] 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, Jan. 4, 2000.

[19] L.-J. 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.

[20] 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, Jan. 2000.

[21] S. G. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989.

[22] 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.

[23] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," in *Readings in Computer Vision*. Amsterdam, The Netherlands: Elsevier, 1987, pp. 671–679.

[24] 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, Jan. 2002.

[25] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[26] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2596–2608, Jun. 2014.

[27] C. Ballester, V. Caselles, L. Igual, J. Verdera, and B. Rougé, "A variational model for P+XS image fusion," *Int. J. Comput. Vis.*, vol. 69, no. 1, pp. 43–58, Aug. 2006.

[28] F. Pérez-Bueno, M. Vega, J. Mateos, R. Molina, and A. K. Katsaggelos, "Variational Bayesian pansharpening with super-Gaussian sparse image priors," *Sensors*, vol. 20, no. 18, p. 5308, Sep. 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/18/5308

[29] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2822–2834, Jul. 2013.

[30] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10257–10266.

[31] Z.-C. Wu, T.-Z. Huang, L.-J. Deng, J. Huang, J. Chanussot, and G. Vivone, "LRTCFPan: Low-rank tensor completion based framework for pansharpening," *IEEE Trans. Image Process.*, vol. 32, pp. 1640–1655, 2023.

[32] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[33] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "A practical compressed sensing-based pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 629–633, Jul. 2012.

[34] M. Ghahremani, Y. Liu, P. Yuen, and A. Behera, "Remote sensing image fusion via compressive sensing," *ISPRS J. Photogramm. Remote Sens.*, vol. 152, pp. 34–48, Jun. 2019.

[35] M. Wang, K. Zhang, X. Pan, and S. Yang, "Sparse tensor neighbor embedding based pan-sharpening via N-way block pursuit," *Knowl.-Based Syst.*, vol. 149, pp. 18–33, Jun. 2018.

[36] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "Two-step sparse coding for the pan-sharpening of remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1792–1805, May 2014.

[37] M. Simsek and E. Polat, "Performance evaluation of pan-sharpening and dictionary learning methods for sparse representation of hyperspectral super-resolution," *Signal, Image Video Process.*, vol. 15, no. 6, pp. 1099–1106, Sep. 2021.

[38] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[39] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–12.

[40] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[41] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[43] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[44] X. Li, F. Xu, X. Lyu, H. Gao, Y. Tong, S. Cai, S. Li, and D. Liu, "Dual attention deep fusion semantic segmentation networks of large-scale satellite remote-sensing images," *Int. J. Remote Sens.*, vol. 42, no. 9, pp. 3583–3610, May 2021.
[45] X. Li, F. Xu, R. Xia, T. Li, Z. Chen, X. Wang, Z. Xu, and X. Lyu, "Encoding contextual information by interlacing transformer and convolution for remote sensing imagery semantic segmentation," *Remote Sens.*, vol. 14, no. 16, p. 4065, Aug. 2022.
[46] 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.
[47] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
[48] T. R. Shaham, T. Dekel, and T. Michaeli, "SinGAN: Learning a generative model from a single natural image," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4569–4579.
[49] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.
[50] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[51] 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.
[52] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "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.
[53] Z.-R. Jin, T.-J. Zhang, T.-X. Jiang, G. Vivone, and L.-J. Deng, "LAGConv: Local-context adaptive convolution kernels with global harmonic bias for pansharpening," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 1, pp. 1113–1121.
[54] H. Liu, L. Deng, Y. Dou, X. Zhong, and Y. Qian, "Pansharpening model of transferable remote sensing images based on feature fusion and attention modules," *Sensors*, vol. 23, no. 6, p. 3275, Mar. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/6/3275
[55] 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. (ICCV)*, Oct. 2017, pp. 1753–1761.
[56] S. Peng, L.-J. Deng, J.-F. Hu, and Y. Zhuo, "Source-adaptive discriminative kernels based network for remote sensing pansharpening," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–7.
[57] M. Zhou, J. Huang, Y. Fang, X. Fu, and A. Liu, "Pan-sharpening with customized transformer and invertible neural network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, 2022, pp. 3553–3561.
[58] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[59] K. Su, D. Yu, Z. Xu, X. Geng, and C. Wang, "Multi-person pose estimation with enhanced channel-wise and spatial information," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5667–5675.
[60] Z. Huang, Y. Ben, G. Luo, P. Cheng, G. Yu, and B. Fu, "Shuffle transformer: Rethinking spatial shuffle for vision transformer," 2021, *arXiv:2106.03650*.
[61] L. Sun, J. Pan, and J. Tang, "ShuffleMixer: An efficient ConvNet for image super-resolution," 2022, *arXiv:2205.15175*.
[62] J. Pang, C. Jiang, Y. Chen, J. Chang, M. Feng, R. Wang, and J. Yao, "3D shuffle-mixer: An efficient context-aware vision learner of transformer-MLP paradigm for dense prediction in medical volume," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1241–1253, May 2023.
[63] 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, Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 1–3.
[64] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Presses des MINES, 2002.
[65] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, Jan. 1998.
[66] G. Vivone, L. Alparone, J. Chanussot, M. D. Mura, A. Garzelli, G. A. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.
[67] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, 2009.
[68] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.
[69] S. Lolli, L. Alparone, A. Garzelli, and G. Vivone, "Haze correction for contrast-based multispectral pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2255–2259, Dec. 2017.

HONGHUI JIANG (Member, IEEE) received the B.S. degree in biomedical engineering and the M.S. degree in control engineering from the Hefei University of Technology, in 2015 and 2018, respectively. From 2018 to 2020, he was with ZTE Communications Company Ltd. He is currently with the Anhui Technical College of Mechanical and Electrical Engineering. He has published three articles. His research interests include remote sensing, image processing, and deep learning.

HU PENG (Member, IEEE) received the B.S. degree in radio engineering from Anhui University, in 1984, and the M.S. degree in circuits and systems and the Ph.D. degree in biomedical engineering from the University of Science and Technology of China, in 1990 and 1997, respectively.

He has published more than a dozen articles and holds more than seven patent applications. His research interests include ultrasound imaging, electrocardiogram modeling and simulation, electroencephalography, and deep learning.

GUOZHENG ZHANG (Member, IEEE) received the B.S. degree in mechanical design, manufacturing, and automation from the Anhui University of Science and Technology, in 2004, and the M.S. degree in mechatronics engineering and the Ph.D. degree in mechanical design, manufacturing, and automation from the Hefei University of Technology, in 2009 and 2018, respectively.

He has published more than ten articles and holds more than ten patent applications. His research interests include modern integrated manufacturing systems, precision CNC gear machining, and deep learning.