IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2020, accepted August 17, 2020, date of publication August 24, 2020, date of current version September 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3019201

Two-Stage Pansharpening Based on Multi-Level Detail Injection Network

JIANWEN HU[1,2], (Member, IEEE), CHENGUANG DU[1], AND SHAOSHENG FAN[1,2]
[1]School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China
[2]Key Laboratory of Electric Power Robot of Hunan Province, Changsha University of Science and Technology, Changsha 410114, China

Corresponding author: Jianwen Hu (hujianwen1@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61601061 and Grant 61971071, in part by Scientific Research Fund of Hunan Provincial Education Department under Grant 14B006, in part by Open Research Fund of Key Laboratory of Electric Power Robot of Hunan Province under Grant PROF1902.

ABSTRACT Pansharpening is an effective technology to obtain high resolution multispectral (HRMS) images by fusing low resolution multispectral (LRMS) images and high resolution panchromatic (PAN) images. With the rapid development of deep learning, some pansharpening methods based on deep learning have been proposed. Although fused images are greatly improved, there are still some areas for improvement. For example, the spectral preservation is not good enough and the details of fused images are not rich enough. To address the above problems, a two-stage pansharpening method based on convolutional neural network (CNN) is proposed. In the first stage, image super-resolution technology with residual block is used to enhance LRMS. In order to preserve spectra, inspired by the SAM (spectral angle mapper) index, a new spectral loss function is proposed. The second stage is the fusion stage. Detail injection block is proposed by combining detail injection and CNN in this stage. Experiments on WorldView2 and GeoEye1 images demonstrate that our fused images present more spatial details and better spectra by comparing with existing methods.

INDEX TERMS Pansharpening, detail injection block, residual learning, convolutional neural network.

I. INTRODUCTION

Remote sensing images are widely used in many fields such as classification and detection. Panchromatic (PAN) images and multispectral (MS) images are acquired simultaneously by some satellites such as WorldView2 (WV2), WorldView3 (WV3), QuickBird (QB), and GeoEye1 (GE1). Due to the limitations of some objective conditions, the PAN image with high spatial resolution contains little spectral information. Although the MS image presents large amounts of spectral information, its spatial resolution is usually only one fourth of the corresponding PAN image. Only PAN image or MS image could not meet practical needs, and high resolution multispectral (HRMS) images are needed. Pansharpening aims to provide HRMS images by fusing low-resolution multispectral (LRMS) images and PAN images [1].

In the past few decades, many pansharpening methods have been proposed. There are four representative categories: component substitution (CS) [2], multi-resolution analysis (MRA) [3], sparse representation and deep learning. CS-based methods first transform MS image into another space, which can separate the spatial structure and spectral information into different components. Subsequently, the component with spatial structure of transformed MS image is replaced by the PAN image. The classic CS-based methods include intensity-hue-saturation (IHS) fusion method [4], principal component analysis (PCA) fusion method [5], and Gram–Schmidt (GS) fusion method [6]. CS-based methods can obtain rich detail, but the spectral distortion is usually serious. The core of MRA-based methods is multi-scale detail extraction and injection. In general, the spatial details are firstly extracted from the PAN image by MRA, and then injected into the up-sampled multispectral (UPMS) images. The widely used MRA methods include the Laplacian pyramid [8], wavelet transform [9]–[11], curvelet transform [12], non-subsampled contourlet transform [13], [14], sheartlet transform [15], and non-subsampled sheartlet transform (NSST) [42]. Compared with CS-based methods, MRA-based methods present better spectra. To combine the advantages of different pansharpening methods, some hybrid approaches [16], [17], [49] are proposed. In [49], Kwan *et al*. proposed a fusion strategy for

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

WV3 satellite images and a new no-reference image index GQNR by combining the remote sensing image index (D_λ) and the natural image quality index (NIQE).

In the past few years, sparse representation has drawn significant research interest [32]. The core idea of sparse representation is that image can be represented as linear combination of the fewest atoms in an over-complete dictionary. Some pansharpening methods based on sparse representation were proposed in [32]–[36]. Ayas *et al*. took texture information into account in the fusion process, which protects spectra and details better [35]. Gogineni *et al*. proposed a multi-scale learned dictionary for high frequency component [36]. Although the pansharpening methods based on sparse representation achieve good performance, they are usually time consuming.

In recent years, remote sensing fusion methods based on convolutional neural network (CNN) received lots of attention. Some CNN-based pansharpening methods have been proposed, e.g., pansharpening by convolutional neural networks (PNN) [19], Target-PNN [20], multi-scale and multi-depth network (MSDCNN) [21], deep network for pansharpening (PanNet) [22], remote sensing image fusion with deep convolutional neural network (RSIFNN) [23], convolutional autoencoder-based MS fusion (CAE) [41]. In [48], CNN is used to estimate the degradation blur kernel of MS images, which improve the adaptivity of pansharpening method. Although MS/Hyperspectral (HS) image fusion is a relatively new topic in remote sensing, a number of relevant literatures have been published. In [43], 3-D CNN was used to fuse MS and HS images. Two branches network was proposed in [44]. Before fusing MS and HS images, two branches are used to extract spectral and spatial information from HS and MS images, respectively. Han *et al*. proposed a HS and MS image fusion method by combining cluster and multi-branch neural networks [50]. Super-resolution and hybrid color mapping were combined to fuse a high-resolution color image and a low-resolution HS image in [51]. Compared with traditional CS-based and MRA-based algorithms, the CNN-based methods significantly improve the pansharpening performance. There are still some problems in these methods. For example, both Target-PNN and RSIFNN lack of specific detail processing, which results in that the details of fused images are not sharp enough. Although PanNet sharpens spatial details, the relationship among spectral channels of MS image is not considered, which may result in some spectral distortion.

In order to preserve spectra and enrich details of fused images, we propose a two-stage pansharpening method with a new spectral loss function based on the following three motivations.

1. In general, the UPMS images are directly used to fuse with PAN images in the pansharpening methods based on CNN. However, this way does not make full use of UPMS images, which may result in spectral distortion. Our method includes super-resolution stage and fusion stage. The super-resolution stage with residual block is used to enhance the spatial resolution of UPMS images and preserve spectra. In fusion stage, multi-level detail injection network is proposed to further enhance the spatial details of super-resolution MS images.

2. The idea of detail injection was used in traditional methods and got good performance. We combine CNN and the idea of detail injection.

3. MSE (mean square error) is commonly used as the loss function between the super-resolution image and reference image. But MSE is a pixel-wise loss and lacks the relationship among spectral bands, which lead to spectral distortion. In order to reduce spectral distortion, we propose a new spectral loss function inspired by SAM index.

The remainder of this paper is organized as follows. Section II is related work, which describes detail injection, pansharpening and super-resolution with residual learning. The proposed method is presented in Section III. Section IV gives the experimental results and analysis. Finally, Section V gives conclusion and future work.

II. RELATED WORK

A. DETAIL INJECTION FOR PANSHARPENING

In traditional methods [1], the details of PAN image can be injected UPMS image as follows:

$$\begin{aligned} \widehat{\mathbf{MS}}_k &= \widetilde{\mathbf{MS}}_k + g_k \times \mathbf{P}_{dt} \quad k = 1, 2, \cdots, K \\ \mathbf{P}_{dt} &= \mathbf{P} - \mathbf{P}_L, \end{aligned} \tag{1}$$

where $\widetilde{\mathbf{MS}}_k$ and $\widehat{\mathbf{MS}}_k$ represents the k-th band of the UPMS image and the fused image, respectively, K is the number of bands, $\mathbf{P}$ denotes the PAN image. $\mathbf{P}_L$ and $\mathbf{P}_{dt}$ represent the approximation and detail part of PAN image, respectively, and g_k is the injection weight.

According to the formula (1), pansharpening can be decomposed into the following steps. First, the appropriate $\mathbf{P}_{dt}$ should be obtained. Its spatial resolution is the same as that of HRMS image. In general, the approximation part $\mathbf{P}_L$ is obtained by low-pass filtering, and $\mathbf{P}_{dt}$ is created by subtracting the low-pass approximation of PAN image from PAN image [37]. However there exist obvious differences between the details of PAN and MS images, because the spectral range of PAN images and each band of MS images is different. In order to get required detail image, we need to multiply $\mathbf{P}_{dt}$ by a weight g_k, which influences the spectral and spatial quality of fused images. Therefore, $\mathbf{P}_{dt}$ and g_k are important for generating excellent pansharpened images. Some detail injection-based methods have been proposed. BDSD [18] is representative injection algorithm. Liu *et al*. proposed locally linear detail injection method [38], which is based on the assumption that the spatial details of each band of MS image can be locally and linearly represented by the spatial details of PAN images. In [39], the PAN image is decomposed into a low-frequency layer, an edge layer, and a detail layer. The edge layer and the detail layer are injected into the MS image by a proportional injection model. In [40], the spatial details are first extracted from the MS

and PAN images. Then the details are sparsely represented. In order to refine joint details, they designed an adaptive weight factor. Finally, the refined joint details are injected into the MS image by modulation weight to get the fusion result. Inspired by the idea of detail injection, multi-level detail injection network is proposed to achieve image fusion in this paper.

B. PANSHARPENING AND SUPER-RESOLUTION WITH RESIDUAL LEARNING

It is well known that ResNet [30] proposed by He *et al.* is very effective. Its core idea is to form residual through an identity mapping, which can transmit information to the next level well and reduce the difficulty of network learning. The network with residual learning can converge quickly. It has good performance when the complexity of network structure is increased. Some ResNet-based methods have been proposed for the pansharpening problem. The first work using residual learning for pansharpening is the deep residual pansharpening neural network (DRPNN) [52]. Target-PNN [20] is a simple and effective three layers CNN with the idea of residual learning. Researchers tried to use complex structures to design pansharpening network, which make the network having stronger learning ability. Yang *et al.* combined inception module and residual learning to propose a multi-scale and multi-depth network (MSDCNN) [21]. PanNet [22] also used the residual module to build the network model by paying attention on details of fused image, which make the spatial quality of fused image better. According to different characteristics of PAN and MS images, a two-branch network called RSFINN [23] was proposed by extracting features of PAN and MS images respectively. RSIFNN used the idea of residual learning by adding long shortcut.

In recent years, some ResNet-based methods have been proposed for image super-resolution. Long shortcut is used to learn the residual information in [45], [53], which make network converge quickly. Various residual modules have been proposed by combining some technologies and residual learning. Residual dense block (RDB) [54] was proposed by combining dense network and residual connection for image super-resolution. Dense residual generative adversarial network (DRGAN) [56] uses RDB block as basic block to implement remote sensing super-resolution. Attention mechanism and residual block are combined for image super-resolution in [46], [47]. Residual channel attention [57] is used in remote sensing super-resolution. Due to the powerful performance of residual learning, we also use it to design our super-resolution network.

III. PROPOSED METHOD

This section is divided into three subsections. Firstly, the overall framework of proposed method is given. Secondly, the super-resolution stage is described. Thirdly, the fusion stage with multi-level details injection is introduced.

A. OVERALL FRAMEWORK

Generally, the deep learning-based pansharpening methods belong to supervised learning. The learning process can be regarded as the minimization of the following formula:

$$l = \left\| f\left(\widetilde{\mathbf{MS}}, \mathbf{P}, w\right) - \mathbf{X} \right\|_2, \tag{2}$$

where $\mathbf{X}$ is reference image, f and w denote network and related parameters, respectively, l is loss function.

Fig. 1 shows our pansharpening framework, which consists of SR stage and fusion stage. The UPMS, SRMS and HRMS images are the input, output and label images of SR stage, respectively. In fusion stage, SRMS image and the panchromatic detail $\mathbf{P}_{dt}$ are the input, fused MS image is the output, and HRMS image is the label image. Although our approach is a two-stage network with two-stage loss, the two-stage network still is an end-to-end network. In the first stage, super-resolution technology is used to enhance spatial resolution and protect spectra simultaneously. In order to preserve spectra effectively, a new spectral loss function is proposed. In the fusion stage, the details of PAN images are injected into the enhanced MS images. An effective detail injection module is proposed in this stage. Multi-level details are obtained by stacking this module. Fused images with richer detail are obtained by fusing multi-level detail features. Our method can be regarded as the minimization of the following formula:

$$l_{all} = (1 - w_1) \times l_{sr} + w_1 \times l_{fusion}, \tag{3}$$

where l_{sr}, l_{fusion} and l_{all} represent super-resolution loss, fusion loss and the total loss, respectively, w_1 represents the balance parameter.

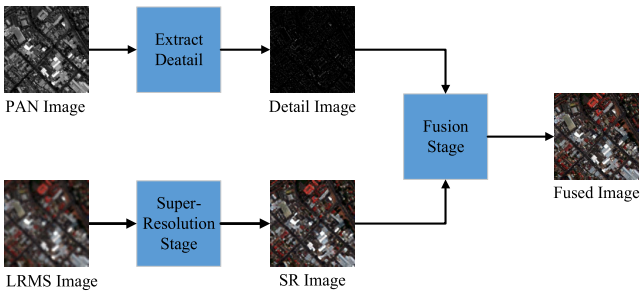


FIGURE 1. Schematic of the proposed two-stage pansharpening method.

B. SUPER-RESOLUTION STAGE

Generally, the input of existing CNN-based methods is PAN and UPMS images. In this way, the UPMS image is not utilized effectively. Fusion result depending on PAN images excessively may lead to spectral distortion. In this paper, we fully utilize UPMS image to preserve spectra and improve the spatial resolution by super-resolution technology.

As shown in Fig 2, the super-resolution stage is composed of feature extraction, non-linear mapping and reconstruction. First, low resolution features are extracted by a convolutional

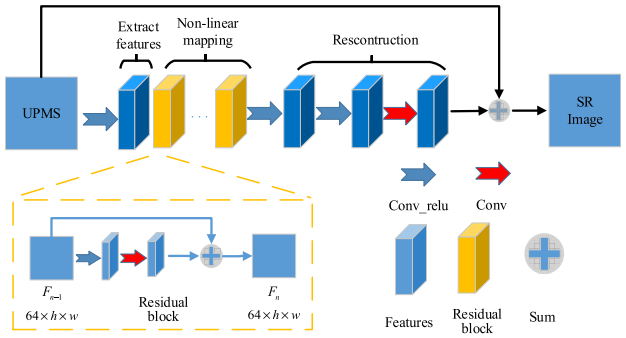


FIGURE 2. Network structure of the super-resolution stage.

layer. It can be expressed by the following formula:

$$\begin{aligned} Fea\_LR &= S_1(\widetilde{\mathbf{MS}}) \\ S_1(A) &= Relu\left(Conv_1\left(A\right)\right) \\ Conv_i(A) &= w_i * A + b_i \\ Relu\left(A\right) &= \max\left(0, A\right), \end{aligned} \tag{4}$$

where $S_1(\cdot)$ represents low-resolution feature extraction. Fea_LR denotes low-resolution features. $Conv_i(\cdot)$ is the i-th convolution, $Relu(\cdot)$ is the ReLU (rectified linear unit) activation function.

Then, high resolution features can be obtained by non-linear mapping. Residual blocks are used to implement non-linear mapping. The core idea of residual block is to form residual through an identity mapping, which can transmit information to the next level well and reduce the difficulty of network learning. Non-linear mapping can be expressed as:

$$\begin{aligned} Fea\_HR &= S_2\left(Fea\_LR\right) \\ S_2\left(A\right) &= R_m\left(R_{m-1}\cdots\left(R_1\left(A\right)\right)\right) \\ R_i\left(A\right) &= Conv_{2i+1}\left(Relu\left(Conv_{2i}\left(A\right)\right)\right) + A, \end{aligned} \tag{5}$$

where $S_2(\cdot)$ represents non-linear mapping, Fea_HR denotes high resolution features. $R_i(\cdot)$ is the i-th residual block, m denotes the number of residual blocks.

Finally, super-resolution (SR) images are reconstructed from high resolution features. Long shortcut as a residual connection is added in the super-resolution network so that the network converges quickly and the efficiency is higher. It can be expressed by

$$\begin{aligned} Pre\_SR &= S_3\left(Fea\_HR\right) + \widetilde{\mathbf{MS}} \\ S_3\left(A\right) &= Conv_{2n+3}\left(Relu\left(Conv_{2n+2}\left(A\right)\right)\right), \end{aligned} \tag{6}$$

where $S_3(\cdot)$ represents reconstruction, Pre_SR denotes the reconstructed super-resolution MS (SRMS) image.

MSE (mean square error) is commonly used as the loss function between the reconstructed super-resolution image and reference image. But MSE is a pixel-wise loss and lacks the relationship among spectral bands, which lead to spectral distortion. The SAM metric calculates the angle between the corresponding pixels of the fused image and reference image, which can quantify the spectral distortion. The SAM is defined as:

$$SAM\left(\mathbf{I}, \mathbf{J}\right) = arccos\left(\frac{\langle \mathbf{I}, \mathbf{J}\rangle}{\|\mathbf{I}\|_2 \|\mathbf{J}\|_2}\right) \tag{7}$$

where $\mathbf{I}$ and $\mathbf{J}$ are pixel vector with size $1 \times K$, and K is the number of bands in MS images.

To facilitate solving gradient and back propagation, we use the absolute value function to substitute arccos function after calculating the spectral correlation between SRMS image and HRMS image. The highest value of correlation is 1, and the higher correlation means better fusion performance. In order to be consistent with minimization optimization, we subtract the correlation from 1. The proposed spectral loss l_{spe} is defined as follows:

$$\begin{aligned} &l_{spe}(Pre\_SR, X) \\ &= \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\left|\frac{\langle Pre\_SR\left(i,j\right), X\left(i,j\right)\rangle}{\|Pre\_SR\left(i,j\right)\|_2 \|X\left(i,j\right)\|_2} - 1\right| \end{aligned} \tag{8}$$

where Pre_SR and X are SRMS image and reference image, respectively. $\langle\cdot,\cdot\rangle$ denotes the inner product. Both $Pre\_SR\left(i,j\right)$ and $X\left(i,j\right)$ are a vector with size $1 \times K$.

The loss of the super-resolution stage is weighted average of MSE loss and spectral loss, i.e.,

$$l_{sr} = w_2 \times \|Pre\_SR - X\|_2 + (1 - w_2) \times l_{spe}\left(Pre\_SR, X\right), \tag{9}$$

where the weight w_2 is used to balance two kinds of losses.

C. FUSION STAGE WITH MULTI-LEVEL DETAIL INJECTION

LRMS images lack of details, and PAN images contain a lot of high-resolution details. Inspired by detail injection idea, we propose detail injection block by combining CNN and detail injection. The proposed multi-level detail injection network is shown in Fig 3. SR image and $\mathbf{P}_{dt}$ are used as the input of fusion stage.

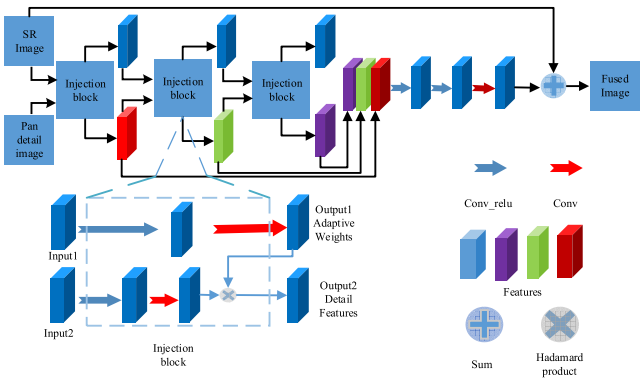


FIGURE 3. Fusion stage with multi-level detail injection.

As previously mentioned in section 2.1, first we need to obtain the suitable details $\mathbf{P}_{dt}$ of PAN image. The PAN image is filtered by the mean filter with size 5×5 to get $\mathbf{P}_L$, then the PAN details $\mathbf{P}_{dt}$ is obtained by subtracting $\mathbf{P}_L$ from the PAN

image. However, $\mathbf{P}_{dt}$ is a single detail map, which is difficult to satisfy the requirement. It is feasible to extract multiple detail features from $\mathbf{P}_{dt}$ by CNN for its strong learning and non-linear representation ability. Generally, the spatial details of PAN and MS images are different, and corresponding injection weight for each band can adjust the injected details to avoid some artifacts. So, the second step is to find the appropriate weight. Each band of MS image has its own characteristics, and pansharpening is to obtain HRMS image. Therefore, the weights should be obtained from an image that is similar to HRMS image. Compared with UPMS image, SRMS image is more similar to HRMS image. Therefore, we obtain the weights by extracting the features of SRMS image. Finally, the Hadamard product is used to get the detail features of fused image. In order to obtain more details, we stack injection block to extract multi-level detail features. Multi-level detail injection can be expressed by

$$\begin{aligned} \mathbf{d_{multi}} &= C\left(\mathbf{d_1}, \mathbf{d_2}, \cdots, \mathbf{d_P}\right) \\ \mathbf{w_p} &= F_p^w\left(\mathbf{w_{p-1}}\right), \quad \mathbf{d_p} = F_p^d\left(\mathbf{d_{p-1}}\right) \otimes \mathbf{w_p}, \\ & \qquad p = 1, 2, \ldots, P \\ F_p^w\left(\mathbf{w_{p-1}}\right) &= Conv\left(Relu\left(Conv\left(\mathbf{w_{p-1}}\right)\right)\right) \\ F_p^d\left(\mathbf{d_{p-1}}\right) &= Conv\left(Relu\left(Conv\left(\mathbf{d_{p-1}}\right)\right)\right), \end{aligned} \tag{10}$$

where $F_p^d\left(w_{p-1}\right)$ and $F_p^w\left(d_{p-1}\right)$ represent the p-th level detail features extraction network and adaptive weights extraction network, respectively. $\otimes$ denotes Hadamard product. w_p is the p-th adaptive weight, and d_p is the obtained p-th detail features. w_0 and d_0 are SRMS image and the details of PAN image, respectively. They are the input of the first level injection block, and are also the input of the fusion stage. d_{multi} represents multi-level detail features. C denotes the concatenation of multi-level detail features on channel dimension. P is the number of proposed injection block.

The final details are obtained by fusing multi-level detail features. We add long shortcut, and the details are added into the SRMS image. In this stage, it can be regarded as the minimization of the following model:

$$\begin{aligned} l_{fusion} &= \|Pre\_Fusion - X\|_2 \\ Pre\_Fusion &= F_{fusion}\left(d_{multi}\right) + Pre\_SR \\ F_{fusion}\left(d_{multi}\right) &= Conv_3\left(Relu\left(Conv_2\left(Relu\left(Conv_1\left(d_{multi}\right)\right)\right)\right)\right), \end{aligned} \tag{11}$$

where F_{fusion} denotes the multi-level detail features fusion which is built by three simple convolution layers. Pre_Fusion, Pre_SR and X denote fused image, SRMS image and reference image, respectively.

IV. EXPERIMENTAL RESULTS AND ANALYSES

In this section, we performed some experiments on GeoEye-1 (GE1) and WorldView-2 (WV2) images. Their spatial resolution and band information are shown in Tables 1 and 2. First, the ablation experiments and parameter selection are given to analyze the network structure. Then our method is compared with four traditional methods (SFIM [17], MTF-GLP-HPM [8], BDSD [18] and ATWT [9]) and three CNN-based methods (Target-PNN [20], RSIFNN [23] and PanNet [22]). Six indices (Q [25], SAM [26], ERGAS [24], SCC [28], Q4 [27], Q_{2n} [31]) are used to evaluate the quality of fused images at the reduced scale. Q evaluates the structure similarity between fused images and reference images. Q4 is the vector extension of the Q index. Q_{2n} is suitable for the assessment of images with the number of spectral bands greater than four. Spectral angle mapper (SAM) represents spectral distortion by calculating the average angle between the corresponding spectral vector of fused image and reference image. Relative global dimensional synthesis error (ERGAS) reflects image comprehensiveness distortion. Spatial correlation coefficient (SCC) reflects the correlation between HRMS image details and fused image details. Four commonly used indices (D_λ [29], D_s [29], QNR [29], and SAM) are used to evaluate the quality of fused images at the full scale. D_λ and D_s reflect spectral distortion and loss of spatial detail, respectively. QNR is comprehensiveness distortion through combining D_λ and D_s.

TABLE 1. Spatial resolutions for GeoEye-1 and WorldView-2 sensors.

Sensor	PAN	MS
GE1	0.46 m	1.84 m
WV2	0.46 m	1.84 m

TABLE 2. Spectral wavelength range (in nm) of GeoEye-1 and WorldView-2 sensors.

	Pan	Coastal	Blue	Green	Yellow
GE1	450-900	no	450-520	520-600	no
WV2	450-800	400-450	450-510	510-580	585-625
	Red	Red edge	Nir1	Nir2	
GE1	625-695	no	760-900	no	
WV2	630-690	705-745	770-895	860-1040	

The training images are generated according to Wald's protocol. We rotate the data sets 90 degree, 180 degree and 270 degree, and extract 9801 samples on WV2 and 12,800 samples on GE1 as training set. The patch size is 64 × 64, and the batch size is 64. The test images include 26 images on GE1 and 55 images on WV2. We use TensorFlow framework to implement the proposed method and select the Adam optimizer. The initialization method is Xavier uniform initializer. Long shortcut and local residual connection are used in our method, which make the network converge quickly. We do not utilize any tricks such as gradient clipping to deal with gradient vanishing or explosion. Network parameter setting is given in Table 3.

A. PARAMETER SELECTION AND NETWORK STRUCTURE ANALYSIS

The loss of super-resolution stage is composed of MSE loss and spectral loss. MSE loss focuses on optimizing the spatial

TABLE 3. Network parameter setting, K denotes the number of bands of MS image.

	Super-resolution stage			Fusion stage	
	Extraction	mapping	Restoration	Extraction	Fusion
Layers	1	8	2	6	3
Features	64	64	32-K	32	64-32-K

part of the MS image, while spectral loss function preserves the spectral part of the MS image. We use the parameter w_2 to balance the relationship between them. We study the influence of the parameter w_2 on fused image quality. The experimental results are given in Table 4. ERGAS is a comprehensive image quality index. We analyze the image quality through using the index ERGAS. When w_2 is 1, the fusion result is the worst, which shows that our spectral loss function is effective. As the value of w_2 rises from 0.1 to 0.6, the value of ERGAS fluctuates. Thus, the fusion result is sensitive to small w_2. When w_2 is 0.6, the best fusion result is obtained. With the value of w_2 rising to 0.9, the value of ERGAS rises slightly. Although the image quality drops slightly, the image quality is high, which shows that the fusion result is not sensitive to large w_2. Therefore, the proposed spectral loss function is effective, and the combination of MSE and our spectral loss function can improve the image quality. We set w_2 to 0.6 according to the above results.

TABLE 4. Influence of w_2 on fused image quality.

w_1 / w_2	Q	SAM	ERGAS	SCC	Q4
0.6/0.1	0.9513	3.4069	2.7172	0.9152	0.9548
0.6/0.2	0.9530	3.4193	2.6636	0.9191	0.9563
0.6/0.3	0.9497	3.4310	2.7557	0.9112	0.9535
0.6/0.4	0.9521	3.4149	2.6804	0.9162	0.9555
0.6/0.5	0.9518	3.4558	2.7038	0.9137	0.9554
0.6/0.6	**0.9553**	**3.3808**	**2.5710**	**0.9246**	**0.9586**
0.6/0.7	0.9537	3.4239	2.6464	0.9196	0.9571
0.6/0.8	0.9534	3.4142	2.6538	0.9192	0.9568
0.6/0.9	0.9530	3.4475	2.6653	0.9191	0.9565
0.6/1.0	0.9485	3.5558	2.8137	0.9061	0.9522
Ideal	1	0	0	1	1

In addition, the total loss is composed of super-resolution loss and fusion loss. We use the parameter w_1 to balance the relationship between them. We study the influence of the parameter w_1 on fused image quality. The experimental results are given in Table 5. When w_1 equals 1, the fusion result is the worst, which shows that our two-stage loss is effective. As the value of w_1 rises from 0.1 to 0.9, the value of ERGAS only rises or drops slightly. The image quality is high and stable, which shows that the fusion result is not sensitive to w_1. When w_1 is 0.6, the best fusion result is obtained. Thus, the parameter w_1 is set to 0.6.

TABLE 5. Influence of w_1 on fused image quality.

w_1 / w_2	Q	SAM	ERGAS	SCC	Q4
0.1/0.6	0.9536	3.4139	2.6394	0.9201	0.9569
0.2/0.6	0.9533	3.4040	2.6570	0.9191	0.9566
0.3/0.6	0.9547	3.3948	2.6036	0.9233	0.9580
0.4/0.6	0.9544	3.3945	2.6075	0.9226	0.9576
0.5/0.6	0.9551	**3.3764**	2.5957	0.9242	0.9584
0.6/0.6	**0.9553**	3.3808	**2.5710**	**0.9246**	**0.9586**
0.7/0.6	0.9549	3.3982	2.5983	0.9231	0.9582
0.8/0.6	0.9544	3.3874	2.6106	0.9225	0.9578
0.9/0.6	0.9551	3.3941	2.5960	0.9232	0.9583
1.0/0.6	0.9503	3.4049	2.7555	0.9094	0.9540
Ideal	1	0	0	1	1

Our super-resolution network is mainly composed of residual module (RM). To verify the effectiveness of the super- resolution network, residual module (RM) used in our super-resolution network is compared with standard convolution module (SCM) [45], residual channel attention module (RCAM) [47] and residual attention module (RAM) [46]. The experimental results are shown in Table 6. Compared with other modules, RM presents better results.

TABLE 6. Influence of different modules in super-resolution stage on fusion results.

Module	Q	SAM	ERGAS	SCC	Q4
RAM	0.9495	3.4518	2.7825	0.9077	0.9532
RCAM	0.9515	3.4431	2.7196	0.9151	0.9549
SCM	0.9529	3.4102	2.6622	0.9191	0.9563
RM	**0.9553**	**3.3808**	**2.5710**	**0.9246**	**0.9586**
Ideal	1	0	0	1	1

The proposed detail injection block consists of two branches. Adaptive weights are obtained from the first branch, and the second branch is to generate the detail feature maps. In order to verify the effectiveness of this module, we performed some comparative experiments. First, we study the necessity of two branches by comparing single branch (Fig. 4a) with our two branches (Fig. 4c). Second, the way to generate the weights is important. Our three-dimensional weights are directly obtained by CNN. Squeeze and excitation block (SE-block) [55] (Fig. 4b) is widely used to obtain channel attention weights. We study the influence of different way of weight generation on fusion performance by comparing SE-block (Fig. 4b) with ours (Fig. 4c). The experimental results over GE1 and WV2 dataset are shown in Tables 7 and 8, respectively. The comparison between single branch and our two branches shows that two branches are necessary. Moreover, our method gives better result than SE-block, which means that our weights are more appropriate. The fusion result with our detail injection block presents the best performance. Therefore, the proposed detail injection bock is effective.

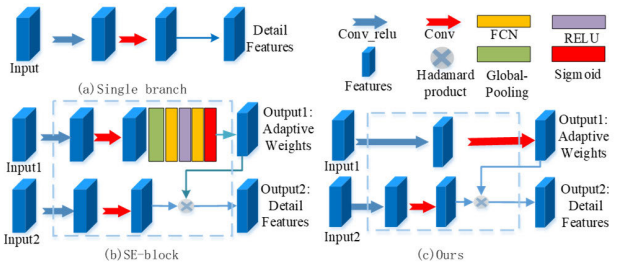


FIGURE 4. Different modules for fusion stage. (a) Single branch. (b) SE-block (c) Ours.

TABLE 7. Influence of different module in fusion stage on GE1 dataset.

	Q	SAM	ERGAS	SCC	Q8
Single	0.9389	3.8284	3.1315	0.8825	0.9433
SE-block	0.9391	3.8053	3.1219	0.8838	0.9436
Ours	**0.9553**	**3.3808**	**2.5710**	**0.9246**	**0.9586**
ideal	1	0	0	1	1

TABLE 8. Influence of different module in fusion stage on WV2 dataset.

	Q	SAM	ERGAS	SCC	Q4
Single	0.9314	4.9628	2.9953	0.9206	0.9372
SE-block	0.9320	4.9490	2.9935	0.9218	0.9376
Ours	**0.9474**	**4.2515**	**2.5385**	**0.9484**	**0.9525**
Ideal	1	0	0	1	1

B. ABLATION EXPERIMENTS

Our method consists of two stages, i.e., super-resolution stage and fusion stage. Super-resolution stage is used to preserve spectra and enhance spatial resolution simultaneously. Fusion stage with multi-level detail injection network generate richer details. We analyze their impact on fusion results by comparing the following seven cases.

1. We give the performance of up-sampled multispectral (UPMS) image obtained by bicubic interpolation.

2. The network only includes super-resolution (SR) stage and does not include fusion stage.

3. Firstly, UPMS images are super-resolved by our SR network to obtain super-resolution multispectral (SRMS) images. Then, SRMS images are fused with PAN images by guided filtering (GF) (the GFCS-B method in [58]).

4. The network includes fusion stage with single-level (SL) detail injection and does not include super-resolution stage.

5. The network includes fusion stage with multi-level (ML) detail injection and does not include super-resolution stage.

6. The network includes fusion stage with single-level detail injection and super-resolution stage (SLSR).

7. The network includes fusion stage with multi-level detail injection network and super-resolution stage (MLSR).

The evaluation indices on WV2 and GE1 dataset are given in Tables 9 and 10, respectively. The best performance is obtained by the MLSR, which proves that our method is effective. Comparing SR with bicubic interpolation, the SAM index is decreased by 1.9 and 1.7 on GE1 and WV2 dataset, respectively, and the SCC index is improved by 0.17 and 0.19 on GE1 and WV2, respectively. Therefore, the super-resolution stage can effectively improve image quality. Although the spatial quality of UPMS image has been improved, it is not enough. The spatial resolution ratio between PAN image and MS image from the same satellite is usually 4. It is difficult to improve the resolution of image by 4 times. The fusion stage is used to further improve the image quality of SRMS. Comparing SR with MLSR, the SAM index is decreased by 0.7 and 1.2 on GE1 and WV2 dataset, respectively, and the SCC index is improved by 0.19 and 0.24 on GE1 and WV2 dataset, respectively. It is obvious that MLSR is much better than SR. Therefore, fusion stage can further improve fusion performance, and PAN image provides important contribution to fusion result.

TABLE 9. Average indices of 26 test images from GE1 on ablation experiment.

	SR	Fusion	Q	SAM	ERGAS	SCC	Q4
Bicubic			0.6609	6.0272	6.6057	0.5626	0.6833
SR	√		0.8510	4.0963	4.7767	0.7349	0.8611
SRGF	√	GF	0.9020	4.0144	4.4536	0.8240	0.9088
SL		Single	0.9263	4.4380	3.3819	0.8573	0.9320
SLSR	√	Single	0.9390	3.7108	3.0963	0.8812	0.9434
ML		Multi	0.9413	3.6212	3.0077	0.8889	0.9457
MLSR	√	Multi	**0.9553**	**3.3808**	**2.5710**	**0.9246**	**0.9586**
Ideal			1	0	0	1	1

TABLE 10. Average indices of 55 test images from WV2 on ablation experiment.

	SR	Fusion	Q	SAM	ERGAS	SCC	Q8
Bicubic			0.6217	7.1822	6.8594	0.5075	0.6527
SR	√		0.7934	5.4612	5.3827	0.7019	0.8107
SRGF	√	GF	0.8910	5.4490	4.3382	0.8548	0.8994
SL		Single	0.9327	5.0361	2.9163	0.9256	0.9392
SLSR	√	Single	0.9413	4.5191	2.7121	0.9394	0.9468
ML		Multi	0.9452	4.3334	2.5852	0.9458	0.9508
MLSR	√	Multi	**0.9474**	**4.2515**	**2.5385**	**0.9484**	**0.9525**
Ideal			1	0	0	1	1

Comparing the SL with the SLSR, the value of SAM on GE1 and WV2 dataset decreases by 0.72 and 0.52, respectively. Therefore, SR stage can get better spectral preservation. Compared with SRGF, SLSR presents better fusion performance, which demonstrates that the proposed injection block is effective. Moreover, the spatial quality of multi-level injection is better than that of single-level injection. SCC is improved significantly. It demonstrates that multi-level detail injection can get richer detail than single detail injection. Therefore, SR stage can effectively preserve spectra, and multi-level detail injection can provide richer details. The fusion quality will be further improved through combining the two stages.

C. EXPERIMENTS AT REDUCED SCALE

The mean value of evaluation indices of fused images on GE1 and WV2 is given in Tables 11 and 12, respectively. The method with the best performance among other methods is compared with our method in the following analyses.

From Table 11, it can be seen that our fusion result gives the best performance. SAM is decreased by 0.48 compared with PanNet, which indicates that our fusion result presents better spectra. SCC is increased by 0.04 compared with PanNet. Thus, more details are injected into fused images of our method. From Table 12, it can be seen that the SAM of our method is decreased by 0.37 compared with Target-PNN. It proves that SR stage can effectively protect spectra. SCC is increased by 0.01 compared with PanNet, which shows that detail injection network generates more abundant details.

TABLE 11. Average indices of 26 test images from GE1 at reduced scale.

Method	Q	SAM	ERGAS	SCC	Q4
MTF-GLP-HPM	0.8593	5.7470	4.4537	0.7876	0.8714
SFIM	0.8235	5.8042	4.9077	0.7761	0.8369
ATWT	0.8470	5.7993	4.6224	0.7817	0.8610
BDSD	0.8686	6.5910	4.6560	0.7884	0.8893
RSIFNN	0.9303	4.2871	3.2913	0.8667	0.9357
Target-PNN	0.9355	3.8993	3.1852	0.8711	0.9403
PanNet	0.9378	3.8668	3.1110	0.8826	0.9425
Ours	**0.9553**	**3.3808**	**2.5710**	**0.9246**	**0.9586**
Ideal	1	0	0	1	1

TABLE 12. Average indices of 55 test images from WV2 at reduced scale.

Method	Q	SAM	ERGAS	SCC	Q8
MTF-GLP-HPM	0.8583	6.8839	4.2291	0.8533	0.8750
SFIM	0.8401	6.9214	4.5091	0.8496	0.8585
ATWT	0.8516	6.8265	4.2997	0.8440	0.8609
BDSD	0.8311	9.0307	5.0683	0.8166	0.8388
RSIFNN	0.9361	5.0159	2.8284	0.9310	0.9423
Target-PNN	0.9416	4.6034	2.7220	0.9367	0.9473
PanNet	0.9413	4.6236	2.7169	0.9386	0.9468
Ours	**0.9471**	**4.2515**	**2.5385**	**0.9484**	**0.9525**
Ideal	1	0	0	1	1

A representative fusion result is given for each satellite. First, the fusion results are compared on GE1. The RGB bands are displayed in Fig. 5. From this figure, it can be seen that CNN-based methods present better fused images than traditional methods. In order to give more obvious difference, the absolute value of the difference between fused images and HRMS image is given in Fig. 6. Our fusion result gives less spatial and spectra information loss, especially in the red rectangle area.

Then the fusion performance is analyzed on WV2. The RGB bands of fused images and residual images are presented in Figs. 7 and 8, respectively. Traditional methods still lose some details. All CNN-based methods perform well that can be observed from Fig. 7. Compared with other CNN-based methods, our result displays less error

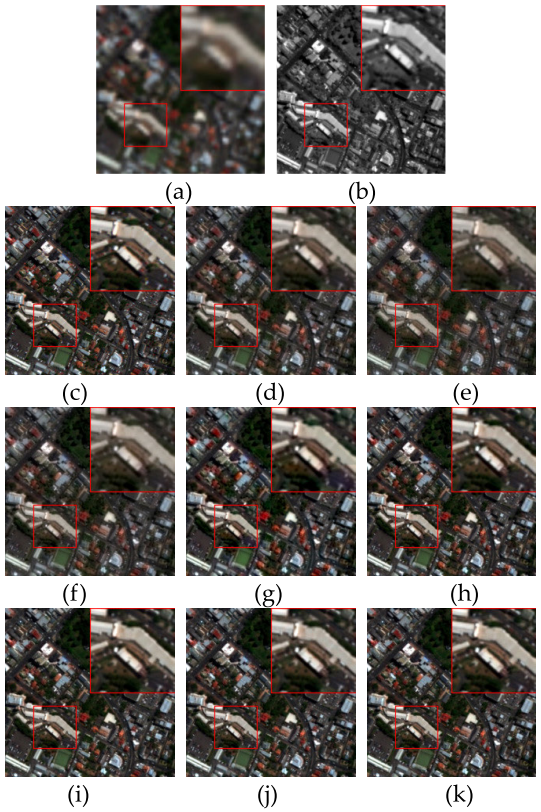


FIGURE 5. Fusion results of the reduced scale on GE1, size: 240 × 240, color channel: RGB. (a) LRMS. (b) PAN. (c) HRMS. (d) MTF-GLP-HPM. (e) SFIM. (f) ATWT. (g) BDSD. (h) RSIFNN. (i) Target-PNN. (j) PanNet. (k) Ours.

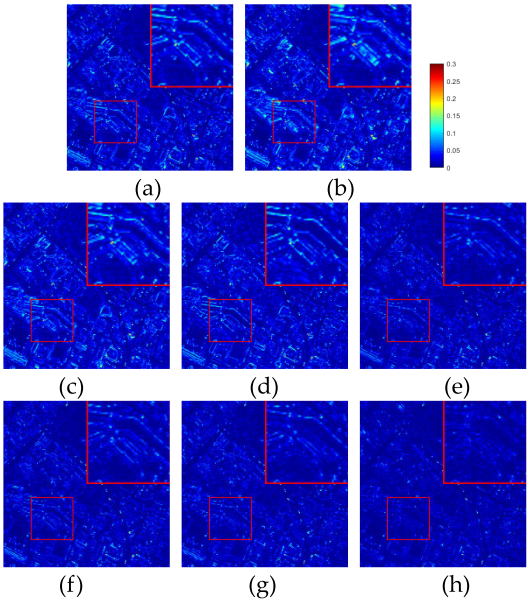


FIGURE 6. Residual images (absolute value after subtraction of HRMS and fusion results) for Figure 5. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

in the red rectangle that can be observed from Fig. 8. Thus, our fusion result shows better spectra and richer details.

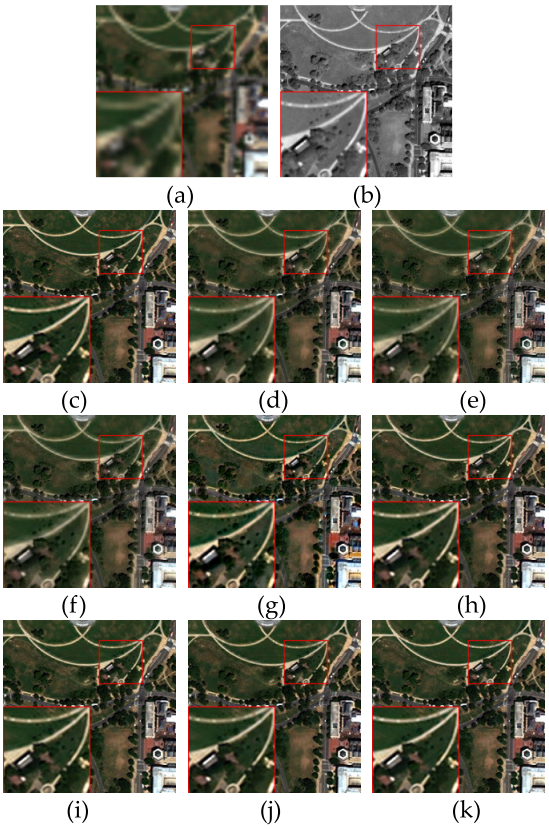


FIGURE 7. Fusion results of the reduced scale on WV2, size: 240 × 240, color channel: RGB. (a) LRMS. (b) PAN. (c) HRMS. (d) MTF-GLP-HPM. (e) SFIM. (f) ATWT. (g) BDSD. (h) RSIFNN. (i) Target-PNN. (j) PanNet. (k) Ours.

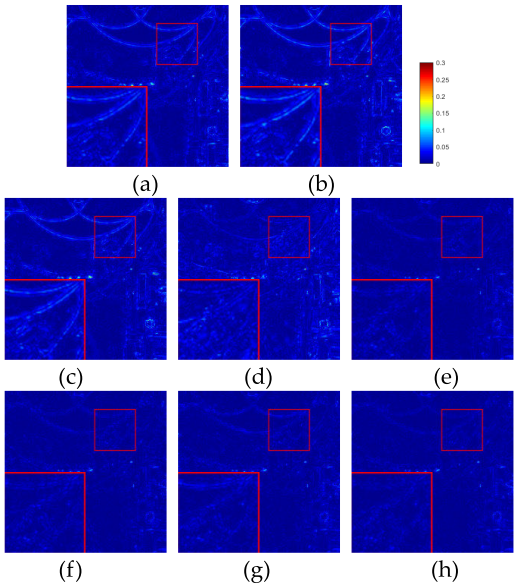


FIGURE 8. Residual images (absolute value after subtraction of HRMS and fusion results) for Figure 7. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

If the response range of PAN image does not cover the spectral range of MS bands, pansharpening task is more difficult. From Table 2, it can be seen that the spectral range

TABLE 13. Average indices of 55 test images from the Costal, NIR1 and NIR2 band of WV2 at reduced scale.

	Q	ERGAS	SAM	SCC	Q8
MTF-GLP-HPM	0.7878	2.8724	5.6136	0.7724	0.7979
SFIM	0.7701	2.8817	5.7524	0.7729	0.7823
ATWT	0.7822	2.8883	5.6776	0.7595	0.7939
BDSD	0.7028	3.1919	7.6296	0.6613	0.7140
RSIFNN	0.9264	2.7447	3.7159	0.8945	0.9294
T-PNN	0.9324	2.6694	3.5489	0.9045	0.9350
PanNet	0.9358	2.6878	3.4625	0.9093	0.9381
Ours	**0.9494**	**2.6365**	**3.1026**	**0.9292**	**0.9514**
Ideal	1	0	0	1	1

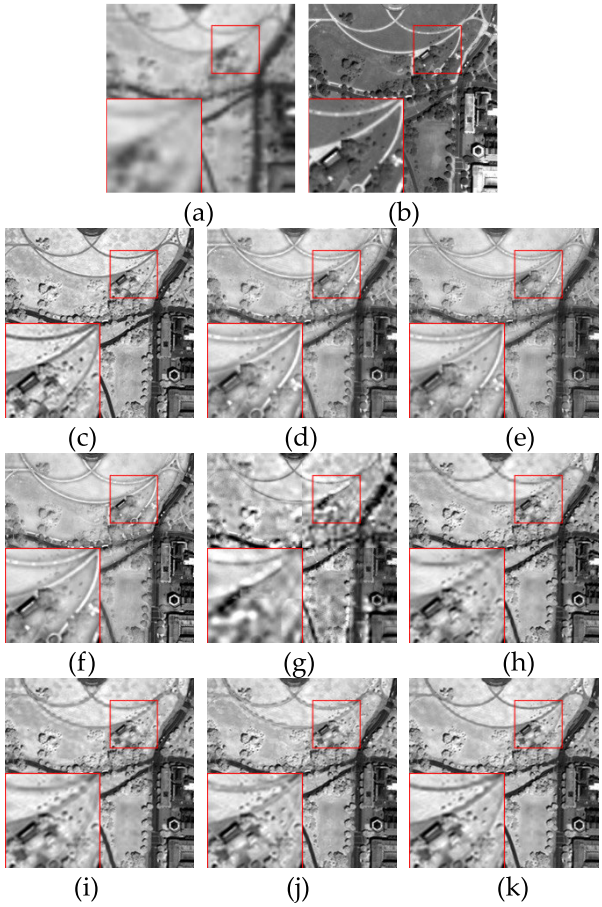


FIGURE 9. Fusion results of the reduced scale on WV2, size: 240 × 240, channel: NIR2. (a) LRMS. (b) PAN. (c) HRMS. (d) MTF-GLP-HPM. (e) SFIM. (f) ATWT. (g) BDSD. (h) RSIFNN. (i) Target-PNN. (j) PanNet. (k) Ours.

of all bands of multispectral image is in the range of PAN image for GE1, while the spectral range of Coastal, NIR1 and NIR2 bands is not in the response range of PAN for WV2. The indexes of three bands (Coastal, NIR1, NIR2) are shown in Table 13. Our method performs the best for all evaluation indexes. The NIR2 band of WV2 images and the corresponding residual images are displayed in Figs. 9 and 10, respectively. It can be seen that our fusion result presents richer detail and less information loss from Figs. 9 and 10.

Therefore, our method preforms the best on both GE1 and WV2. All indices are improved, and the fused images of

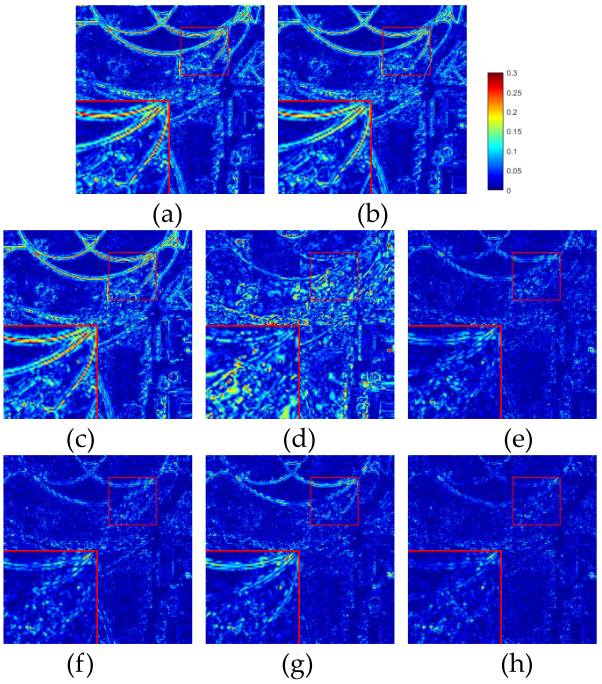


FIGURE 10. Residual images (absolute value after subtraction of HRMS and fusion results) for Figure 9. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

the proposed method are better than that of other methods. It proves the superiority of our method.

D. EXPERIMENTS AT FULL SCALE

In this section, some experiments and analyses at full scale are given. The mean value of evaluation indices on GE1 is given in Table 14. Our fusion result gives the best performance for all evaluation indices. D_λ is decreased by 0.0064 compared with Target-PNN, which shows that our fusion result presents less spectral distortion. D_s is decreased by 0.0055 compared with RSIFNN. It proves that more details are injected by our method. Our result gives the best QNR.

TABLE 14. Average indices of 26 test images from GE1 at full scale.

Method	D_λ	D_s	QNR	SAM
MTF-GLP-HPM	0.1018	0.1132	0.7971	1.8106
SFIM	0.0764	0.1038	0.8282	1.8119
ATWT	0.1014	0.1241	0.7880	1.7933
BDSD	0.0475	0.0271	0.9116	4.4645
RSIFNN	0.0518	0.0272	0.9226	1.2694
Target-PNN	0.0360	0.0351	0.9303	**0.8951**
PanNet	0.0465	0.0280	0.9270	0.9680
Ours	**0.0296**	**0.0217**	**0.9491**	0.9120
Ideal	0	0	1	0

The mean value of evaluation indices on WV2 is given in Table 15. Although our fusion result only gives the second best performance on D_λ and D_s, the SAM and the comprehensive index QNR of our method are the best. SAM is decreased by 0.27 compared with Target-PNN. Although the detail of PanNet is rich, its spectra is not good enough. RSIFNN preserves spectra well, but the details of fusion result are not good enough. Therefore, on the whole, our method gives the best results on both GE1 and WV2.

TABLE 15. Average indices of 55 test images from WV2 at full scale.

Method	D_λ	D_s	QNR	SAM
MTF-GLP-HPM	0.0782	0.1221	0.8094	2.6978
SFIM	0.0625	0.1150	0.8298	2.6995
ATWT	0.0812	0.1276	0.8018	2.5924
BDSD	0.0781	0.1708	0.7638	6.2670
RSIFNN	**0.0204**	0.0594	0.9214	2.5068
Target-PNN	0.0316	0.0504	0.9195	2.2323
PanNet	0.0775	**0.0366**	0.8889	2.3218
Ours	0.0213	0.0368	**0.9426**	**1.9619**
Ideal	0	0	1	0

In the visual comparison part, a pair of source images and theirs fused images at full scale are presented. The RGB bands of the fused images on GE1 are shown in Fig. 11. It can

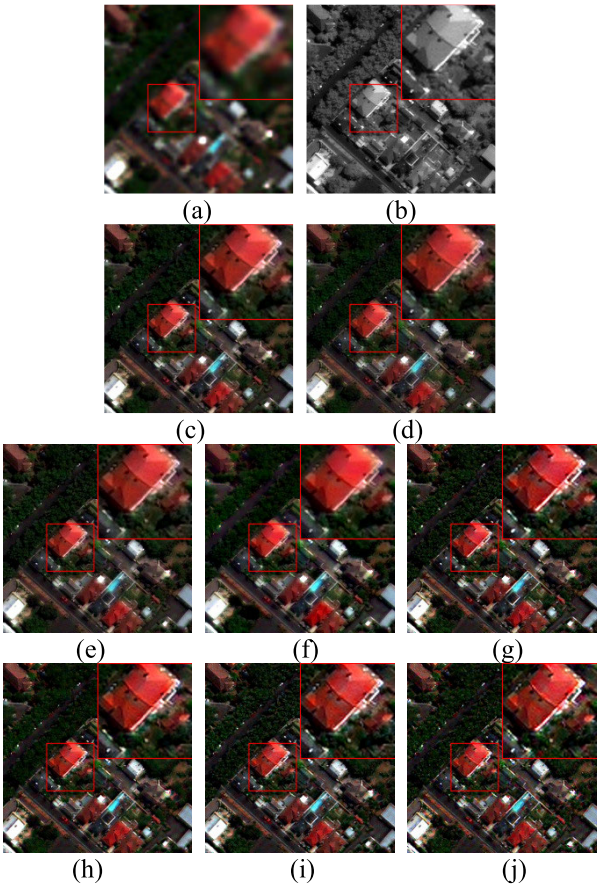


FIGURE 11. Fusion results of the full scale on GE1, size: 240 × 240, color channel: RGB. (a) LRMS. (b) PAN. (c) MTF-GLP-HPM. (d) SFIM. (e) ATWT. (f) BDSD. (g) RSIFNN. (h) Target-PNN. (i) PanNet. (j) Ours.

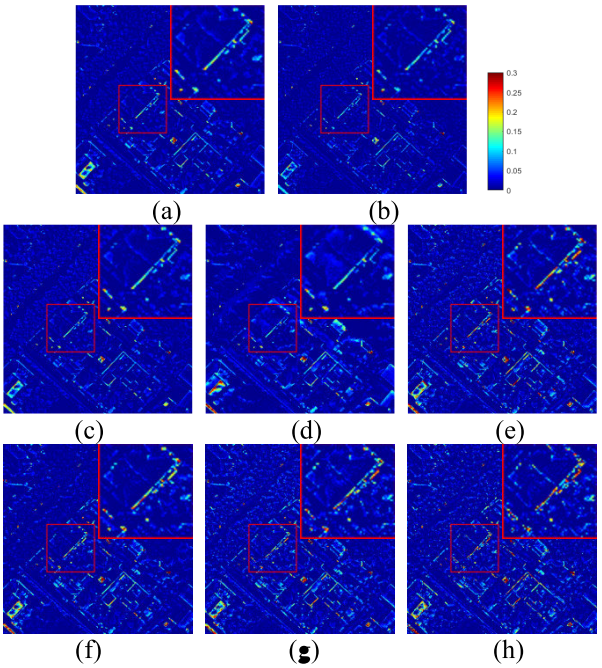


FIGURE 12. Injection detail for Figure 11. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

TABLE 16. Average indices of 55 test images from the Costal, NIR1 and NIR2 band of WV2 at full scale.

	D_λ	D_S	QNR	SAM
MTF-GLP-HPM	0.0754	0.1258	0.8083	2.2310
SFIM	0.0571	0.1101	0.8391	2.2372
ATWT	0.0690	0.1236	0.8160	2.1725
BDSD	0.0707	0.0632	0.8715	5.6294
RSIFNN	0.0277	0.0385	0.9349	2.1829
T-PNN	0.0351	**0.0231**	0.9426	1.9561
PanNet	0.0808	0.0376	0.8854	1.9524
Ours	**0.0199**	0.0260	**0.9546**	**1.6035**
Ideal	0	0	1	0

be observed that the fusion results of CNN-based methods are better than that of the traditional methods.

There is no reference image for full scale. Because the UPMS images as input are sharpened by various methods, the difference between the fusion results and UPMS images can display the injected details and spectral enhancement region. Fig. 12 shows the difference between the RGB bands of fused images and that of UPMS images. Compared with other methods, the proposed method injects more edge details. Compared with the results obtained by PanNet, our fusion results display more spectral enhancement regions, especially in the red rectangle region, which indicates that our method better protects spectral information.

The RGB bands and the difference between fused images and UPMS images on wv2 are given in Figs. 13 and 14, respectively. Our method presents better visual performance than other methods that can be observed from Figs. 13 and 14. Our fusion result shows better spectra and richer details. The evaluation index of three bands (Coastal, NIR1, NIR2)

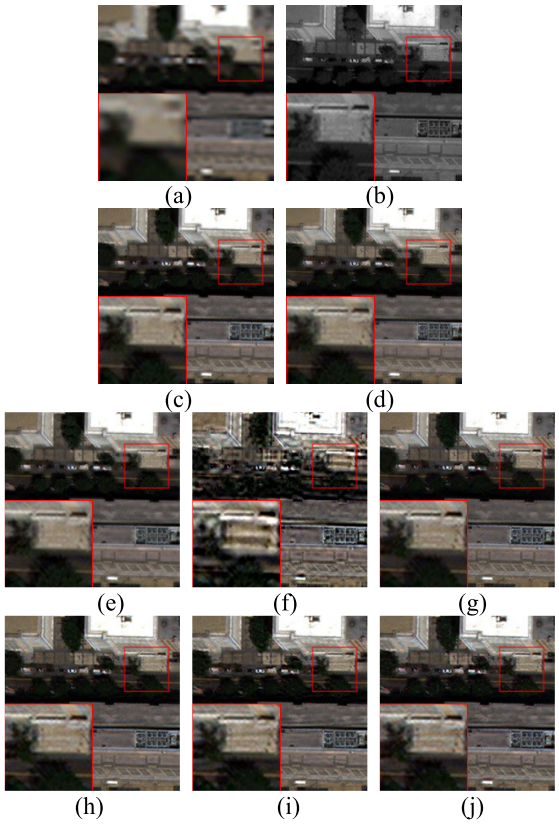


FIGURE 13. Fusion results of the full scale on WV2, size: 240 × 240, color channel: RGB. (a) LRMS. (b) PAN. (c) MTF-GLP-HPM. (d) SFIM. (e) ATWT. (f) BDSD. (g) RSIFNN. (h) Target-PNN. (i) PanNet. (j) Ours.

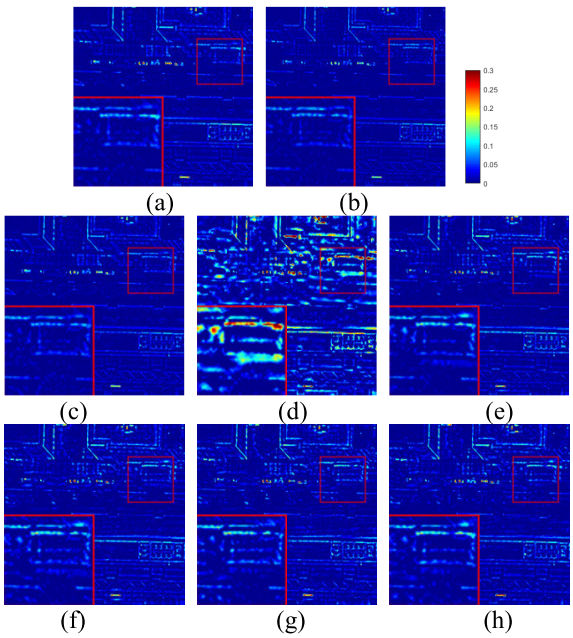


FIGURE 14. Injection detail for Figure 13. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

are shown in Table 16. Although the D_S index is not the best, others indexes are the best. The NIR2 bands and the difference between fused images and UPMS images are

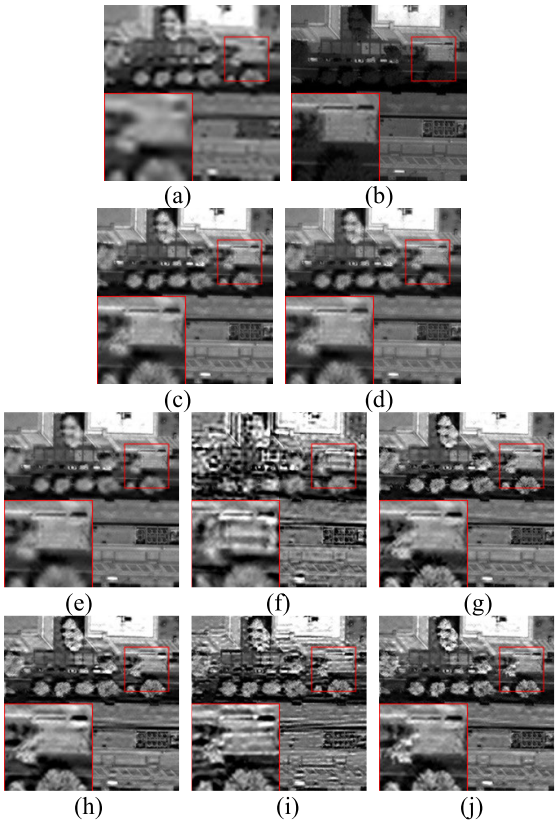


FIGURE 15. Fusion results of the full scale on WV2, size: 240 × 240, channel: NIR2. (a) LRMS. (b) PAN. (c) MTF-GLP-HPM. (d) SFIM. (e) ATWT. (f) BDSD. (g) RSIFNN. (h) Target-PNN. (i) PanNet. (j) Ours.

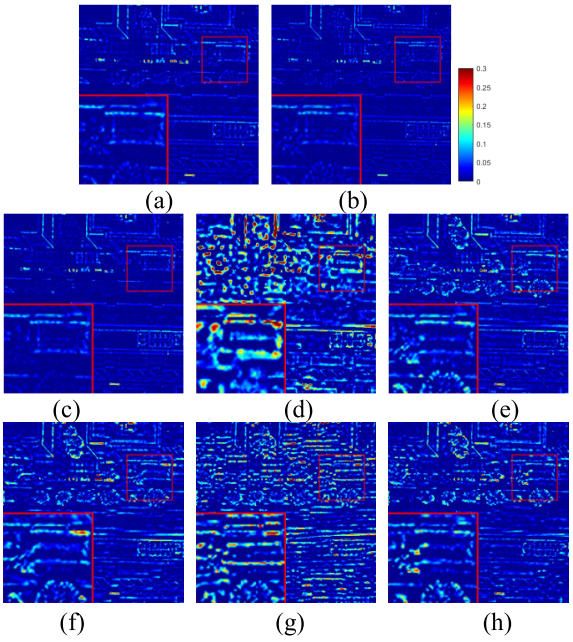


FIGURE 16. Injection detail for Figure 15. (a) MTF-GLP-HPM. (b) SFIM. (c) ATWT. (d) BDSD. (e) RSIFNN. (f) Target-PNN. (g) PanNet. (h) Ours.

given in Figs. 15 and 16, respectively. By observing Fig. 15, the fusion result of PanNet presents obvious noise and the results of traditional methods are smooth. It can be seen that more details are injected into UPMS image by our method. From Fig. 16, it is clear that our fusion result shows richer details, especially in red rectangle area. Therefore, our method improves the fusion results in terms of subjective visual performance and objective indices at full scale.

The computation time of different methods is listed in Table 17 for fusing the UPMS image with size of 480 × 480 × 8 and the PAN image with size of 480 × 480. The traditional methods were measured on CPU, and the CNN based methods were measured on GPU. From this table, it can be seen that the computation time of the traditional methods is less than the CNN based methods. Our method takes more computation time, but 0.2542 second is still acceptable. The training time of CNN based methods is given in Table 18. Although our method needs the longest training time, it only takes 1.1 hours.

TABLE 17. Computation time of different methods.

Machine	i7-7800X CPU 3.50GHz			
Method	SFIM	GLP	ATWT	BDSD
Time(s)	0.0120	0.0380	0.1190	0.1060
Machine	GTX 1080 Ti GPU			
Method	T-PNN	PANNET	RSIFNN	OURS
Time(s)	0.1644	0.1838	0.1934	0.2542

TABLE 18. Training time of different methods.

Methods	PNN	PanNet	Rsifnn	Our
Time(h)	0.52	0.61	0.83	1.10

V. DISCUSSION

Some networks designed for super-resolution task are similar to our proposed network, but there are still many differences among them. Both the networks in [45], [53] and our super-resolution network use long shortcut to learn the residual information, which make network converge quickly. The difference between [45], [53] and ours is that we add local residual block in non-linear mapping part, which further improve non-linear mapping ability. Standard residual modules that contain two convolution layers and an identity mapping are used in our super-resolution network. Compared with standard residual block, residual dense block (RDB) [54] contains many parameters, and its calculation cost is high. Standard residual modules are light and effective. In the part. A, section IV, some experimental results are given. Compared with the network block in [45]–[47], the standard residual module gets better fusion results.

There are some similarities among RAM [46], RCAM [47] and the proposed detail injection block, but they have essential differences. Attention mechanism is used in RAM and RCAM to recognize where or which feature map is important. Then the network focus on optimizing these areas to improve image quality. Pansharpening is to sharpen the LRMS image through using the high-resolution PAN image. The spatial

details extracted from PAN image can indicate the concerned regions, and thus the attention module is not needed. Because the details of the PAN image and the MS image are not the same, we need different weights. The proposed method obtains the weight from the MS image in the other branch. RAM and RCAM learn the corresponding weight from the extracted features, which are single-branch structure. The RCAM obtained the one-dimensional channel attention by global average pooling, two 1×1 convolution layers and softmax operation. The RAM generates one-dimensional channel attention weight by global variance pooling, two 1×1 convolution layers and the two-dimensional spatial attention weight by channel separation convolution, and then combines them to form the three-dimensional weight, while our three-dimensional weight are directly obtained by two convolution layers. Therefore, there are some important differences among our detail injection module, RCAM and RAM.

In addition, our network consists of two stages. The first stage combines the spectral loss function and super-resolution network to preserve spectra and enhance details, and the second stage integrates the detail injection idea into CNN to achieve spatial sharpening. The advantages of two stages are merged into a whole framework to sharpen the LRMS images.

VI. CONCLUSION

In this paper, we proposed a two-stage pansharpening network, which includes super-resolution stage and fusion stage. In super-resolution stage, we make full use of the spectral information of LRMS images to protect spectra. In fusion stage, detail injection block is proposed. It can extract detail features well. The ablation experiment demonstrates the effectiveness of the two stages. The proposed method is compared with other pansharpening methods on GE1 and WV2 satellite image datasets. The experimental results at the reduced and full scale verify the superiority of the proposed method in terms of subjective visual performance and objective evaluation indices.

Remote sensing images are becoming more and more abundant, and their tasks are also diversified. For example, Sentinel-2 satellite provides remote sensing images with three spatial resolution. Sentinel-2 image fusion is more difficult because the spectral range of different bands is different and the number of spatial resolutions is increased. In the future, we will study how to combine the idea of multi-level detail injection with the task of multi-resolution remote sensing image fusion.

REFERENCES

[1] 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.

[2] B. Xie, H. Zhang, and B. Huang, "Revealing implicit assumptions of the component substitution pansharpening methods," *Remote Sens.*, vol. 9, no. 5, pp. 443–457, May 2017.

[3] Y. Yang, W. Wan, S. Huang, P. Lin, and Y. Que, "A novel pan-sharpening framework based on matting model and multiscale transform," *Remote Sens.*, vol. 9, no. 4, pp. 391–411, 2017.

[4] Y. Yang, W. Wan, S. Huang, F. Yuan, S. Yang, and Y. Que, "Remote sensing image fusion based on adaptive IHS and multiscale guided filter," *IEEE Access*, vol. 4, pp. 4573–4582, Aug. 2016.

[5] W. Dong, S. Xiao, X. Xue, and J. Qu, "An improved hyperspectral pansharpening algorithm based on optimized injection model," *IEEE Access*, vol. 7, pp. 16718–16729, Jan. 2019.

[6] X. Li, Y. Zhang, Y. Gao, and S. Yue, "Using guided filtering to improve gram-Schmidt based pansharpening method for GeoEye-1 satellite images," in *Proc. 4th Int. Conf. Inf. Syst. Comput. Technol.* Atlantis Press, 2016, pp. 33–37.

[7] J. Lee and C. Lee, "Fast and efficient panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 155–163, Jan. 2010.

[8] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, May 2006.

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

[10] L. Alparone, S. Baronti, B. Aiazzi, and A. Garzelli, "Spatial methods for multispectral pansharpening: Multiresolution analysis demystified," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 2563–2576, May 2016.

[11] Y.-Z. Zhang, T.-Z. Huang, L.-J. Deng, J. Huang, X.-L. Zhao, and C.-C. Zheng, "A framelet-based iterative pan-sharpening approach," *Remote Sens.*, vol. 10, no. 4, pp. 622–639, Apr. 2018.

[12] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, Apr. 2007.

[13] X. Lu, J. Zhang, and Y. Zhang, "An improved non-subsampled contourlet transform-based hybrid pan-sharpening algorithm," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3393–3396.

[14] R. James and M. Vadivel, "Improvement of spectral and spatial information using modified WAMM and modified bi-cubic interpolation method in non-subsampled contourlet transform domain," in *Proc. Int. Conf. Circuits, Power Comput. Technol.*, Mar. 2015, pp. 1–5.

[15] Z. Zhang, X. Luo, and X. Wu, "A new pan-sharpening method using statistical model and shearlet transform," *IETE Tech. Rev.*, vol. 31, no. 5, pp. 308–316, Sep. 2014.

[16] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "MTF-based deblurring using a Wiener filter for CS and MRA pansharpening methods," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 6, pp. 2255–2269, Jun. 2016.

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

[18] 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.

[19] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, pp. 594–615, Jul. 2016.

[20] 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.

[21] 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.

[22] 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.*, Oct. 2017, pp. 1753–1761.

[23] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.

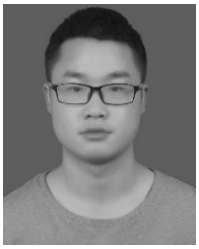
[24] L. Wald, *Data Fusion: Definitions and Architectures: Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses des MINES, 2002.

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

[26] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, 1992, pp. 147–149.

[27] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.
[28] 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.
[29] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[31] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi/hyper-spectral images," *IEEE Geosci. Remote Sens. Lett*, vol. 6, no. 4, pp. 662–665, Oct. 2009.
[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] H. Yin, "PAN-guided cross-resolution projection for local adaptive sparse representation-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4938–4950, Jul. 2019.
[34] H. Wu, S. Zhao, J. Zhang, and C. Lu, "Remote sensing image sharpening by integrating multispectral image super-resolution and convolutional sparse representation fusion," *IEEE Access*, vol. 7, pp. 46562–46574, Apr. 2019.
[35] S. Ayas, E. T. Gormus, and M. Ekinci, "An efficient pan sharpening via texture based dictionary learning and sparse representation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 7, pp. 2448–2460, Jul. 2018.
[36] R. Gogineni and A. Chaturvedi, "Sparsity inspired pan-sharpening technique using multi-scale learned dictionary," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 360–372, Dec. 2018.
[37] A. Garzelli, B. Aiazzi, L. Alparone, S. Lolli, and G. Vivone, "Multispectral pansharpening with radiative transfer-based detail-injection modeling for preserving changes in vegetation cover," *Remote Sens.*, vol. 10, no. 8, p. 1308, Aug. 2018.
[38] J. Liu, Y. Hui, and P. Zan, "Locally linear detail injection for pansharpening," *IEEE Access*, vol. 5, pp. 9728–9738, Jun. 2017.
[39] X. Meng, J. Li, H. Shen, L. Zhang, and H. Zhang, "Pansharpening with a guided filter based on three-layer decomposition," *Sensors*, vol. 16, no. 7, p. 1068, Jul. 2016.
[40] Y. Yang, L. Wu, S. Huang, W. Wan, and Y. Que, "Remote sensing image fusion based on adaptively weighted joint detail injection," *IEEE Access*, vol. 6, pp. 6849–6864, Jan. 2018.
[41] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.
[42] X. Wang, S. Bai, Z. Li, R. Song, and J. Tao, "The PAN and MS image pansharpening algorithm based on adaptive neural network and sparse representation in the NSST domain," *IEEE Access*, vol. 7, pp. 52508–52521, Apr. 2019.
[43] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-Convolutional neural network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.
[44] J. Yang, Y. Q. Zhao, and J. C. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.*, vol. 10, no. 5, pp. 800–822, May 2018.
[45] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[46] J.-H. Kim, J.-H. Choi, M. Cheon, and J.-S. Lee, "RAM: Residual attention module for single image super-resolution," 2018, *arXiv:1811.12043*. [Online]. Available: http://arxiv.org/abs/1811.12043
[47] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.
[48] J. Hu, Z. He, and J. Wu, "Deep self-learning network for adaptive pansharpening," *Remote Sens.*, vol. 11, no. 20, pp. 2395-1–2395-22, Oct. 2019.
[49] 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.
[50] X. Han, J. Yu, J. Luo, and W. Sun, "Hyperspectral and multispectral image fusion using cluster-based Multi-Branch BP neural networks," *Remote Sens.*, vol. 11, no. 10, pp. 1173-1–1173-12, May 2019.
[51] C. Kwan, J. H. Choi, S. H. Chan, J. Zhou, and B. Budavari, "A super-resolution and fusion approach to enhancing hyperspectral images," *Remote Sens.*, vol. 10, no. 9, pp. 1416-1–1416-28, Sep. 2018.
[52] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.
[53] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.
[54] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 2472–2481.
[55] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[56] W. Ma, Z. Pan, F. Yuan, and B. Lei, "Super-resolution of remote sensing images via a dense residual generative adversarial network," *Remote Sens.*, vol. 11, no. 21, pp. 2578-1–2578-24, Nov. 2019.
[57] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, and A. Plaza, "Remote sensing image superresolution using deep residual channel attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9277–9289, Nov. 2019.
[58] J. Liu and S. Liang, "Pan-sharpening using a guided filter," *Int. J. Remote Sens.*, vol. 37, no. 8, pp. 1777–1800, Apr. 2016.

JIANWEN HU (Member, IEEE) received the B.S. degree from the Inner Mongolia University of Science and Technology, Baotou, China, in 2008, and the Ph.D. degree from Hunan University, Changsha, China, in 2013. He is currently a Lecturer with the School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha. His research interests include image processing, deep learning, and sparse representation.



CHENGUANG DU received the B.S. degree from the Changsha University of Science and Technology, Changsha, China, in 2017, where he is currently pursuing the M.S. degree with the School of Electrical and Information Engineering. His research interests include deep learning and remote sensing image fusion.

SHAOSHENG FAN received the B.S. degree from Southwest Jiaotong University, Chengdu, China, in 1987, and the Ph.D. degree from Hunan University, Changsha, China, in 2005. He is currently a Professor with the School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha. His research interests include image processing and electric robot.