Pan-Sharpening Based on Convolutional Neural Network by Using the Loss Function With No-Reference

Zhangxi Xiong, Qing Guo, *Member, IEEE*, Mingliang Liu, and An Li

***Abstract*—In order to preserve the spatial and spectral information of the original panchromatic and multispectral images, this article designs a loss function suitable for pan-sharpening and a four-layer convolutional neural network that could adequately extract spectral and spatial features from original source images. The major advantage of this study is that the designed loss function does not need the reference fused image, and then the proposed pan-sharpening method does not need to make the simulation data for training. This is the big difference from most existing pan-sharpening methods. Moreover, the loss function takes into account the characteristics of remote sensing images, including the spatial and spectral evaluation indicators. We also add the feature enhancement layer in convolutional neural network, thus, the proposed four-layer network contains feature extraction, feature enhancement, linear mapping and reconstruction. In order to evaluate the effectiveness and universality of the proposed fusion model, we selected thousands of remote sensing images that include different sensors, different times and different land-cover types to make the training dataset. By evaluating the performance on the WorldView-2, Pleiades and Gaofen-1 experimental data, the results show that the proposed method achieves optimal performance in terms of both the subjective visual effect and the object assessment. Furthermore, the codes will be available at https://github.com/Zhangxi-Xiong/pan-sharpening.**

***Index Terms*—Convolutional neural network, deep learning, feature enhancement, loss function, pan-sharpening, remote sensing image fusion.**

I. Introduction

PAN-SHARPENING is using a high spectral and low spatial resolution multi-spectral (MS) image with a high spatial resolution panchromatic (PAN) image to produce a high spectral and high spatial resolution image. Many remote sensing pan-sharpening methods have been proposed in recent years and can be mainly classified into three types: component substitution (CS), multiresolution analysis (MRA), and model-based algorithms (MB).

The CS methods rely on the substitution of a component obtained by means of a spectral transformation of the MS data with the PAN image. Classical CS methods include the intensity-hue-saturation transform [1], the principal component analysis [2], Gram-Schmidt (GS) method [3], the GS adaptive (GSA) approach [4], the band-dependent spatial-detail (BDSD) method [5], [6], Brovey transform [7], etc. Most CS methods achieve good performance in the spatial quality but usually cause the spectral distortion with varying degrees.

The MRA methods are based on the injection of spatial details that are obtained through a multiresolution decomposition of the PAN image into the resampled MS bands. The MRA methods mainly include wavelet transform (WT) [8], discrete WT [9], à trous WT [10], Laplacian pyramid (LP) [11], the additive wavelet LHS method [12], beyond wavelets [13]–[15] (e.g., shearlet) and generalized LP (GLP) based on Gaussian filters matching the modulation transfer function (MTF) [16], [17]. MRA methods preserve the spectral characteristics well but are generally not satisfactory in terms of the spatial enhancement.

The MB methods mainly include the sparse representation (SR) and the deep learning algorithms. SR [18]–[20] methods first learn the spectral dictionary from the low spatial resolution data, then combine the known high spatial resolution data to predict the high spatial resolution and high spectral resolution data. In recent years, deep learning algorithm has been widely used in remote sensing field [21]–[24] and there has been an increasing interest in the deep learning fusion algorithms, such as the pan-sharpening method with deep neural networks [25], the pan-sharpening by learning a deep residual network [26], the CNN-based pan-sharpening method (CNNB) [27], and pan-sharpening by convolutional neural networks (PNN) [28].

The selection of loss function of deep learning based pan-sharping method is a very important research point. Literatures [26]–[32] were adopted mean square error (MSE) as loss function, MSE has simple calculation and good data fitting. However, it has a strong penalty for large error and a low penalty for small error, which will ignore the influence of image content itself. In addition, the partial derivative value is very small when the

Manuscript received August 11, 2020; revised October 28, 2020; accepted November 3, 2020. Date of publication November 16, 2020; date of current version January 6, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61771470 and Grant 41590853; in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences under Grant QYZDY-SSWDQC026; and in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDA19060103. *(Corresponding author: Qing Guo.)*

Zhangxi Xiong is with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China, and also with the Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Harbin 150080, China (e-mail: pairs719@163.com).

Qing Guo and An Li are with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China (e-mail: guoqing@aircas.ac.cn; lian@aircas.ac.cn).

Mingliang Liu is with the Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Harbin 150080, China (e-mail: mll_0608@163.com).

Digital Object Identifier 10.1109/JSTARS.2020.3038057

output probability value is close to 0 or close to 1, which may cause the partial derivative value to almost disappear when the model starts training.

In [33], the mean absolute error (MAE) is used as the loss function. Compared with MSE, MAE has better convergence. Since the loss of MSE is much greater than that of MAE in the position with large error, it will give more weight to outliers, and the model will try its best to reduce the error caused by outliers, thus reducing the overall performance of the model. Therefore, the MAE is more effective when there are more outliers in the training data.

In order to prevent the model from over fitting, the regularization term is added to the loss function, such as l_0 norm penalty, l_1 norm penalty (parameter sparsity penalty), and l_2 norm penalty (weight decay penalty). In [25] and [34], the weight decaying term and the sparsity term are added to the MSE loss function.

The cross-entropy loss function is generally used in the field of classification. It can measure the difference between two different probability distributions in the same random variable. In machine learning, it is expressed as the difference between the real probability distribution and the predicted probability distribution. In [35], a pan-sharpening method based on generative adversarial networks (GANs) is proposed. In the GANs, the cross entropy is used to determine the closeness between the actual output and the expected output, which can be regarded as a binary classification problem.

In the most of the methods mentioned above, there is normally no ground truth image as the reference fused image. Therefore, the training dataset of these methods is often made by the degraded original MS and PAN images according to the Wald protocol [36], and the original MS is used as a reference image. In this article, motivated by the super-resolution convolutional neural network (SRCNN) [37], we build a four-layer CNN architecture that is improved from the three-layered SRCNN architecture. First, we add a feature enhancement layer to enhance complex features. Then, according to the characteristics of remote sensing image, a new loss function including the spectral and spatial evaluation is designed, which will make the convolutional result have high spectral similarity with MS and high spatial similarity with PAN. Next, mass remote sensing images as training sets are used to improve the universality of the proposed CNN model. We carry out experiments on three datasets including Pleiades, GaoFen-1, and WorldView-2 multi-resolution sensors, compare results with a score of well-established fusion techniques, and obtain significant improvements on all three datasets under both full-reference and no-reference performance metrics.

The remainder of this article is organized as follows. Section II describes our proposed framework in detail. Section III offers both qualitative and quantitative analyses through three groups' experimental results. Finally, this article draws conclusions and discusses future work in Section IV.

II. Proposed CNNB With Labels of Original Data

In order to improve the lack of prominent details and the presence of blurred edges of some existing CNNBs, and to solve the existing model learning is not real remote sensing data, the pan-sharpening method based on a CNN model is proposed in this article. In order to solve the existing deep learning fusion model needs the simulated degraded source images and is not based on the real remote sensing data, the new loss function with no-reference is designed. In the following, details about datasets are first provided, then our proposed fusion architecture and the improved CNN-based network with labels of original data are described.

TABLE I
Spatial Resolutions for Gaofen-1, Pleiades, and WorldView-2 Sensors

	MS	PAN
Gaofen-1	8m	2m
Pleiades	2m	0.5m
WorldView-2	1.84m	0.46m

A. Datasets

Three different types of remote sensing images are selected as the experimental datasets including Gaofen-1, Pleiades, and WorldView-2. Table I shows the spatial resolution for Gaofen-1, Pleiades, and WorldView-2 sensors. Table II shows the spectral bands of datasets. When the band is comprised the wavelength range (in nm) is reported.

B. Proposed Pan-Sharpening Model

The proposed basic architecture consists of two main parts: The training stage and the testing stage. The training stage is to learn the super parameters by our architecture in a supervised manner, while the testing stage is to produce the high spatial resolution multispectral (HMS) image through the training stage learned super parameters.

Before introducing the training stage, it is better to introduce how to make the training data and label, our training data and label is different from that made by our predecessors. In [25]–[28], the required HMS image is not available which means cannot find the ground truth. In addition, they often use the MSE as the loss function, which needs the reference fused HMS image. Therefore, they preprocess the training data according to the Wald protocol and get the simulated degraded MS and PAN by using the MTF as the input data of the fusion network, while the original MS is considered as the so-called HMS reference image. In this article, we designed a no-reference loss function including the spectral evaluation and spatial evaluation, so there is no need to preprocess to get the simulated input data, while the original MS and PAN images are as the references.

The training data consists of up-sampled MS and original PAN images. Assuming the original MS is an n-band image, then the original MS is up-sampled to the size of PAN. Finally, the n $+ 1$ bands training data is obtained by concatenating the up-sampled MS and the original PAN. Fig. 1 shows the production process of training data.

The label consists of the original MS and the original PAN. Due to the different sizes between the original MS and the original PAN, we should cut the original PAN image to the size

TABLE II
SPECTRAL BANDS FOR GAOFEN-1, PLEIADES, AND WORLDVIEW-2 SENSORS

	PAN	Coastal	Blue	Green	Yellow	Red	Red Edge	NIR	NIR 2
Gaofen-1	450-900	no	450-520	520-590	no	630-690	no	770-890	no
Pleiades	470-830	no	430-550	500-620	no	590-710	no	740-940	no
WorldView-2	450-800	400-450	450-510	510-580	585-625	630-690	705-745	770-895	860-1040

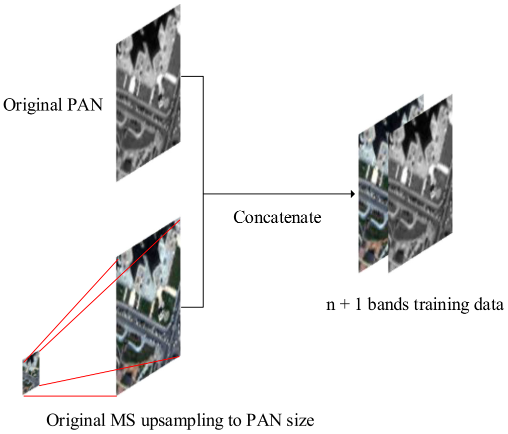

Fig. 1. Production process of training data.

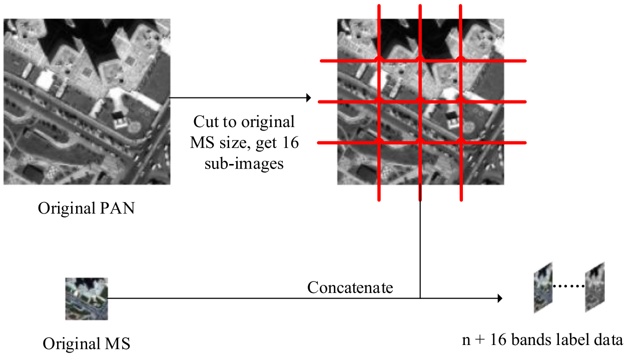

Fig. 2. Production process of label data.

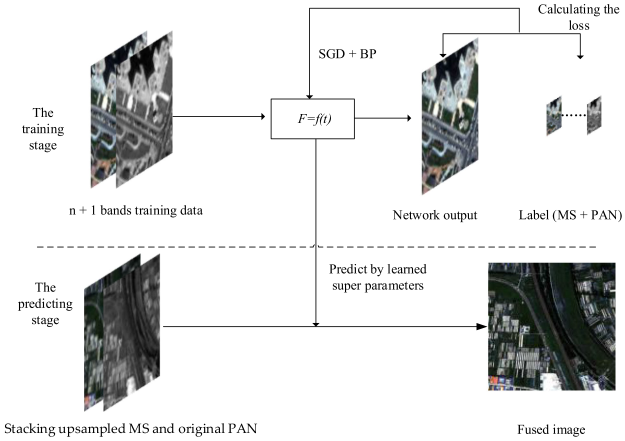

Fig. 3. Workflow of the proposed pan-sharpening method.

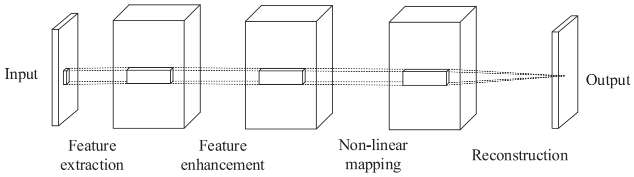

Fig. 4. Structure of pan-sharpening.

of the original MS first. The ratio of PAN to MS size of the three groups satellite images we selected is 4, so PAN will be cut into 16 parts. Finally, the n + 16 bands label is obtained by concatenating the original MS and the cut PAN subimages. Fig. 2 shows the production process of label.

In the training stage, the n + 1 bands training data as the CNN based pan-sharpening network input, the output is the fused image, and the loss function is used to calculate the spectral and spatial losses. In the predicting stage, we choose the same source satellite images that are not in the training data. Then, stacking the up-sampled MS and original PAN as the predicted input. Finally, the fused HMS image is predicted through the training stage learned super parameters. Fig. 3 shows the workflow of the proposed pan-sharpening method.

In Fig. 3, the $F = f(t)$ denotes the pan-sharpening structure, which is based on CNNs. In article [28], the SRCNN has achieved a praiseworthy effect, and it greatly retains the spectral characteristics of the original MS image. However, since its training data is simulated data processed by the Wald protocol and its network structure is simple, the details in PAN cannot be extracted well when using real data to predict. Inspired by the improvement of details enhancement by a feature enhancement layer [38], we try to add a feature enhancement layer to SRCNN and propose a new four-layer CNN pan-sharpening structure.

The proposed structure of pan-sharpening has four parts: feature extraction, feature enhancement, nonlinear mapping, and reconstruction, which is shown in Fig. 4.

1) Feature extraction: using convolution to extract features of the input image, it is similar to map image patch to low resolution dictionary in sparse coding.
2) Feature enhancement: enhancing the feature gotten in the layer of feature extraction.
3) Non-Linear mapping: mapping low-resolution features to high-resolution features, which is similar to find high-

TABLE III
PARAMETERS OF THE PROPOSED CNN BASED PAN-SHARPENING STRUCTURE

c1	K1×K1	f1(x)	c2	K2×K2	f2(x)	c3	K3×K3	f3(x)	c4	K4×K4	f3(x)	c5
n+1	9×9	ReLU	64	7×7	ReLU	32	5×5	ReLU	32	5×5	x	n

resolution dictionary corresponding to image patch in the dictionary learning.

4) Reconstruction: image reconstruction based on high resolution features.

The proposed pan-sharpening structure parameters are shown in Table III, where *n* means the number of the MS bands, c1 means the $n+1$ bands input image which has been described above. The output c5 comprises *n* bands, which corresponds to the MS bands with the resolution of PAN. The $Ki \times Ki$ $(i = 1, 2, 3, 4)$ means the filter kernel size, ci $(i = 2, 3, 4)$ is the number of filters, which corresponds to the output feature maps. Rectified linear unit (ReLU) which function is $\max(0, x)$, is used as the activation function. More details could find in [28] and [38].

During the training stage, stochastic gradient descent algorithm (SGD) and backpropagation are utilized to iteratively learn all of the parameters (w, b) in the network for optimal allocation. Considering the spectral and spatial characteristics of remote sensing images, we design a no-reference loss function to measure the spectral and spatial characteristics of convolutional results and labels at the same time. The spectral distortion index (D_λ) and spatial distortion index (D_s) are employed in the loss function, they are defined as follows: (1) shown at the bottom of this page.

$$D_s \triangleq \sqrt[q]{\frac{1}{L}\sum\nolimits_{l\,=\,1}^{L}\left|Q\left(\text{fused}_l, P\right) - Q\left(MS_l, \tilde{P}\right)\right|^q} \quad (2)$$

where $Q(\text{fused}_l, \text{fused}_r)$ is the quality index values between the l_th band and r_th band of fused image, $Q(MS_l, MS_r)$ calculates the quality index values between the l_th band and r_th band of MS image, $Q(\text{fused}_l, P)$ is the quality index values between the fused image l_th band and PAN, $Q(MS_l, \tilde{P})$ is the quality index values between the MS image l_th band and degraded PAN image. L is the number of MS bands. *p* and *q* are typically set to 1 [39]. D_λ and D_s are always lower than or equal to 1. The closer D_λ and D_s are to 0, the better the evaluation index is. The output is derived from the training data through the pan-sharpening structure, and we name it as $\hat{F}$. Then, when calculating the loss, we divide the tensor label into two parts: MS and PAN. Finally, we calculate the D_λ and D_s respectively, and choose larger as loss function. In this way, D_λ and D_s will gradually tend to zero in thousands of epochs. Fig. 5 shows the workflow of the loss function.

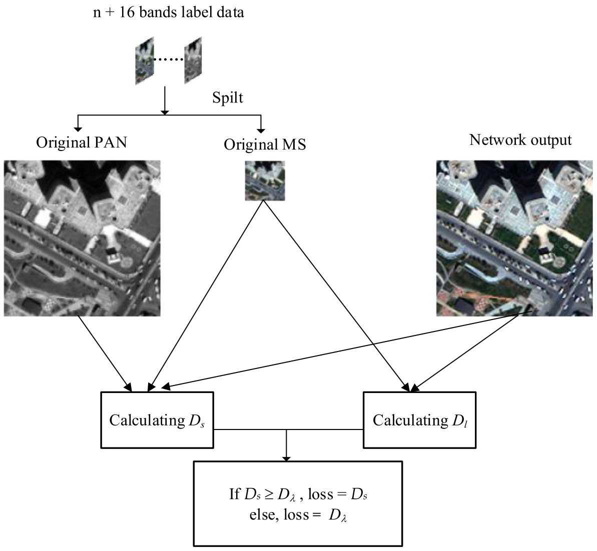

Fig. 5. Workflow of loss function.

The loss function $L(w, b)$ is defined as

$$d_\lambda = \frac{1}{M}\sum_{n=1}^{M} D_\lambda(MS, \widehat{F}) \quad (3)$$

$$d_s = \frac{1}{M}\sum_{n=1}^{M} D_s(MS, PAN, \widehat{F}) \quad (4)$$

$$L(w, b) = \begin{cases} d_\lambda, d_\lambda \geq d_s \\ d_s, d_s > d_\lambda \end{cases} \quad (5)$$

where (w, b) is the set of all involved super parameters of the proposed pan-sharpening architecture which named filter weights and biases, and M represents the number of patches. Using this alternating loss function, the spectral loss and spatial loss of fusion results can be controlled simultaneously. SGD uses a momentum parameter to reduce randomness, therefore

$$(w, b)_{i+1} = (w, b)_i + \Delta(w, b)_i = (w, b)_i + \mu \cdot \Delta(w, b)_{i-1} - \alpha \nabla L_i \quad (6)$$

where μ denotes the momentum and α denotes the learning rate. According to the work of predecessors, the number of iterations has been fixed to 1.12×10^6, $\mu = 0.9$, and $\alpha = 10^{-4}$, except for the last layer, where $\alpha = 10^{-5}$.

$$D_\lambda \triangleq \sqrt[p]{\frac{1}{L\left(L-1\right)}\sum_{l\,=\,1}^{L}\underbrace{\sum_{r\,=\,1}^{L}}_{r\neq 1}\left|Q\left(\text{fused}_l, \text{fused}_r\right) - Q\left(MS_l, MS_r\right)\right|^p} \quad (1)$$

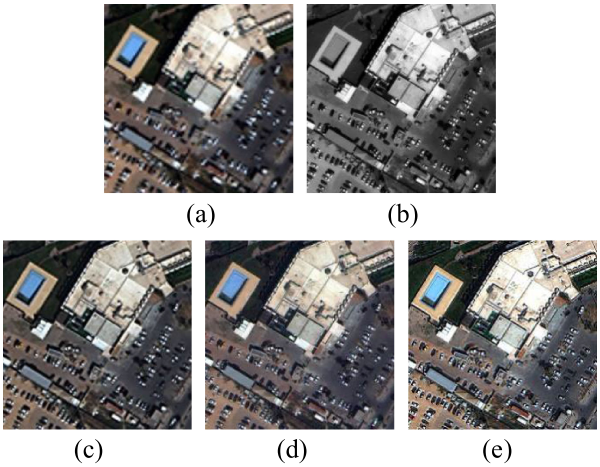

Fig. 6. Comparative experiments. (a) Up-sampled MS. (b) PAN. (c) Output one. (d) Output two. (e) Output three.

TABLE IV
EVALUATION INDEX OF COMPARATIVE EXPERIMENTS

	D_s	D_λ
SRCNN	0.0386	0.1123
Improved SRCNN	0.0287	0.0302
Our model	0.0253	0.0217

In order to verify the feasibility and effectiveness of feature enhancement layer in the improved SRCNN, we designed a comparative experiment. In experiment one, the network is SRCNN and the MSE is used as the loss function. In experiment two, the network is our improved SRCNN. According to the control variable comparison principle, we still use the MSE loss function, not our designed loss function. D_s and D_λ are used as the evaluation index. The test image is from WorldView-2 satellite and the size is 400×400.

In Fig. 6, output one is the SRCNN output, output two is the improved SRCNN output, output three is improved SRCNN using designed loss function. Compared with the up-sampled MS image, we can find that (c)–(e) maintain the good spectral characteristics. However, the underpart of image (c) does not look as normal as image (d), it has a slight spectral distortion, especially around the car. The roof of the building in the SRCNN output looks a little darker, while the roof of the building in the improved SRCNN output looks closer to an up-sampled MS image. This is because some spectral features are not fully preserved and enhanced, while the training data of (d) is enhanced. (e) maintains good spectral information and improves spatial resolution Fig. 6(c) is based on the degraded simulation training data and (e) is based on the original training data. Compared with the original PAN image visually, there is a little bit of difficulty to measure which one has better spatial details, but from Table IV, we can find that the D_s of the improved SRCNN output is less than the SRCNN output, this means that the improved SRCNN has better spatial detail retention ability than the SRCNN. The D_λ of the improved SRCNN output is much less than that of the SRCNN output, which indicates that the spectral retention ability of the improved SRCNN is much better than the SRCNN. After using our designed loss function on the improved SRCNN, D_s and D_λ decrease further.

To sum up, the visual analysis and the evaluation index could prove that our improved SRCNN with no-reference loss function works very useful.

III. Experiments and Analysis

A. Methods for Comparison and Objective Evaluation Metrics

In this article, seven state-of-the-art fusion methods are adopted for comparison, including the additive wavelet luminance proportional (AWLP) [12], the band-dependent spatial-detail with physical constraints (BDSD_PC) [6], the Gram Schmidt adaptive approach (GSA) [4], the MTF-based generalized Laplacian pyramid (MTF_GLP) [17], PanNet: A deep network architecture for pan-sharpening [33], the PNN [28], and the partial replacement adaptive component substitution (PRACS) [40]. For performance assessment, seven widely recognized objective fusion metrics including full-reference (low resolution) and no-reference (full resolution) are applied in our experiments as shown in Table V. where SSIM is defined as

$$\text{SSIM }(F, MS) = \frac{(2\mu_F\mu_{MS} + c_1) * (2\sigma_{FMS} + c_2)}{(\mu_F^2 + \mu_{MS}^2 + c_1) * (\sigma_F^2 + \sigma_{MS}^2 + c_2)} \tag{6}$$

Where μ_F and μ_{MS} denote the mean value of fused image F and MS image, respectively. σ_F^2 and σ_{MS}^2 represents the variance of fused image F and MS image, respectively. The covariance of images is represented by σ_{FMS}, c is a constant. SSIM reflects the structural similarity of two image. The bigger the SSIM, the more similarities between the images. The best value of SSIM is 1.

CC is defined as, (7) shown at bottom of the next page, where $\bar{F}$ and $\overline{MS}$ denote the mean value of fused and MS images, respectively. $CC(F, MS)$ reflects the correlation of the fused and MS images. CC is often used to measure how much spectrum information is preserved. The bigger CC is, the higher the similarity between the two images. The best value of *CC* is 1.

ERGAS is defined as

$$\text{ERGAS} = 100\frac{h}{l}\sqrt{\frac{1}{k}\sum\nolimits_{i\,=\,1}^{k}\left[\frac{\text{RMSE}\,(i)}{\mu\,(i)}\right]^2} \tag{8}$$

$$\text{RMSE }(F, MS) = \sqrt{\frac{1}{M * N}\sum\nolimits_{u=1}^{M}\sum\nolimits_{v=1}^{N}\left[F\,(u, v) - MS\,(u, v)\right]^2} \tag{9}$$

where h and l are the spatial resolutions of the PAN image and MS image, respectively. k is the number of bands of the fused image. RMSE is the root mean squared error. RMSE gives the standard measure of the difference between F and MS. $\mu(i)$ denotes the mean of ith band of reference MS image. The smaller the ERGAS, the better the fusion result. The best value of ERGAS is 0.

TABLE V
FULL-REFERENCE (LOW RESOLUTION) AND NO-REFERENCE (FULL RESOLUTION) PERFORMANCE METRICS

full reference	*SSIM*	Structural similarity index [41]
	CC	Correlation coefficient [42]
	ERGAS	*Erreur Relative Globale Adimensionnelle de Synthèse* [43]
	SAM	Spectral Angle Mapper [44]
no reference	*QNR*	Quality with no-reference index [39,45]
	D_s	Spatial distortion index
	D_λ	Spectral distortion index

TABLE VI
OBJECTIVE EVALUATION OF FUSION RESULTS USING THE FIRST GROUP GAOFEN-1 DATA

	AWLP	BDSD_PC	GS	MTF_GLP	PanNet	PNN	PRACS	Proposed
SSIM	0.6805	0.8509	0.4800	0.6466	0.7569	0.7469	**0.9269**	0.7223
CC	0.9226	**0.9808**	0.8244	0.9156	0.9517	0.9572	0.9603	0.9704
ERGAS	2.1650	1.0715	3.0075	2.2911	1.1092	1.5560	**0.7535**	0.9152
SAM	0.5386	0.4912	1.5330	0.5781	0.5699	0.6576	0.8230	**0.4644**
QNR	0.8560	0.8910	0.6754	0.8300	0.9323	0.9342	0.7005	**0.9500**
D_s	0.0524	0.0516	0.2268	0.0572	0.1003	0.0381	0.2377	**0.0247**
D_λ	0.0967	0.0606	0.1265	0.1197	0.0752	0.0288	0.0811	**0.0259**

SAM is defined as

$$\text{SAM } (v,\hat{v}) = \ arccos\left(\frac{\langle v,\hat{v}\rangle}{\|v\|_2 * \|\hat{v}\|_2}\right). \tag{10}$$

SAM denotes the absolute value of the spectral angle between two vectors. In the formula, v is the original spectral pixel vector, $\hat{v}$ is the distorted vector obtained by applying fusion to the coarser resolution MS data. The zero value of SAM denotes the absence of spectral distortion. SAM is measured in either degree or radian, which is usually averaged over the whole image to yield a global measurement of the spectral distortion.

The quality with no reference (QNR) is defined as

$$\text{QNR } = (1 - D_s)^\alpha * (1 - D_\lambda)^\beta \tag{11}$$

where α and β are the trade off coefficients, usually $\alpha = \beta = 1$. D_s and D_λ are defined in formula (2) and (1). Therefore, the maximum theoretical value of *QNR* is 1 and is obtained when the spatial distortion index D_s and spectral distortion index D_λ are both 0.

B. Experimental Result and Analysis

Figs. 7(a), 8(a), and 9(a) show the up-sampled low-resolution MS images of Gaofen-1, Pleiades, and WorldView-2, respectively. The high-resolution PAN images of Gaofen-1, Pleiades, and WorldView-2 are shown in Figs. 7(b), 8(b), and 9(b). The fusion results of AWLP, BDSD_PC, GS, MTF_GLP, PanNet, PNN, PRACS, and the proposed method are illustrated in Figs. 7(c)–(j), 8(c)–(j), 9(c)–(j), respectively. Tables VI–VIII list those quality indicators of objective evaluations of Figs. 7–9, respectively. All the figures are displayed R (red), G (green), B (blue) bands for natural color composition. All of the objective evaluation metrics are the average values of all test scenes and the best performance is marked in bold font.

1) Result Analysis for Gaofen-1 Data: In Fig. 7, from a visual point of view, the fused image of the GSA approach happens a very serious spectral distortion whose color looks significantly different from that of the MS image. From the local enlarged image, we can see that the spatial details of the fused images of the BDSD_PC method and the PNN method are not rich, it is not as good as that of the AWLP, the MTF_GLP, the PanNet and the proposed methods, and the edge of the PRACS method is not obvious. Compared with the local enlarged image of PAN, there seems to be some noise in the AWLP, the GSA, the MTF_GLP, and the PanNet methods. From Table VI, we can see that our proposed method performs best in term of indicators SAM, QNR, D_s, and D_λ. CC and ERGAS rank second which are only a little worse than the optimal ones.

$$CC\ (F, MS) = \frac{\sum_{u=1}^{M}\sum_{v=1}^{N}\left[F\left(u,v\right) - \bar{F}\right]\left[MS\left(u,v\right) - \overline{MS}\right]}{\sqrt{\sum_{u=1}^{M}\sum_{v=1}^{N}\left[F\left(u,v\right) - \bar{F}\right]^2 * \sum_{u=1}^{M}\sum_{v=1}^{N}\left[MS\left(u,v\right) - \overline{MS}\right]^2}} \tag{7}$$

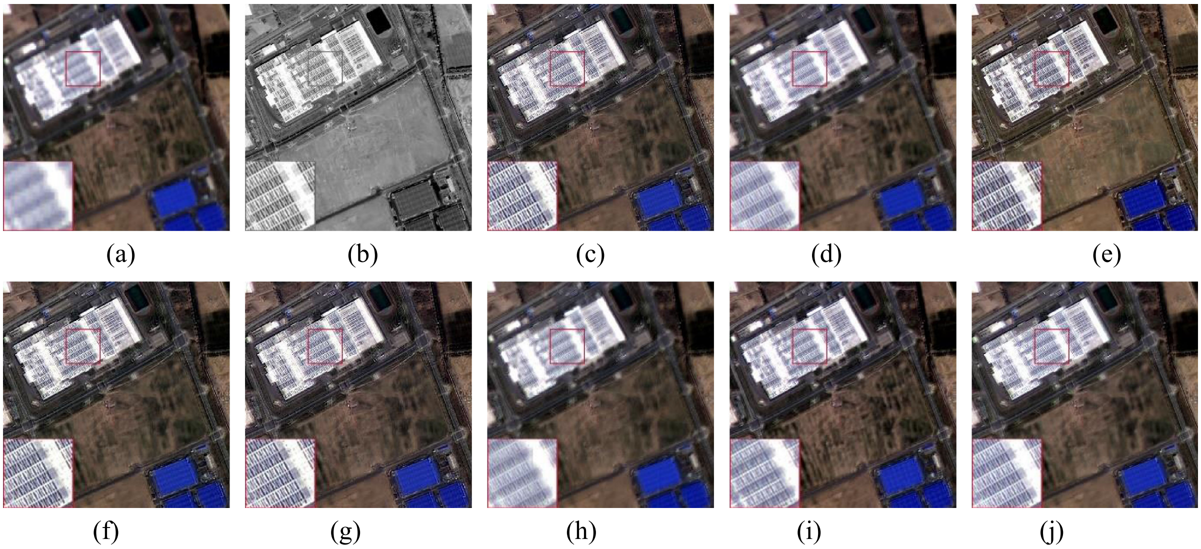

Fig. 7. Fused images by different methods of the first group Gaofen-1 data. (a) MS. (b) PAN. (c) AWLP. (d) BDSD_PC. (e) GSA. (f) MTF_GLP. (g) PanNet. (h) PNN. (i) PRACS. (j) Proposed.

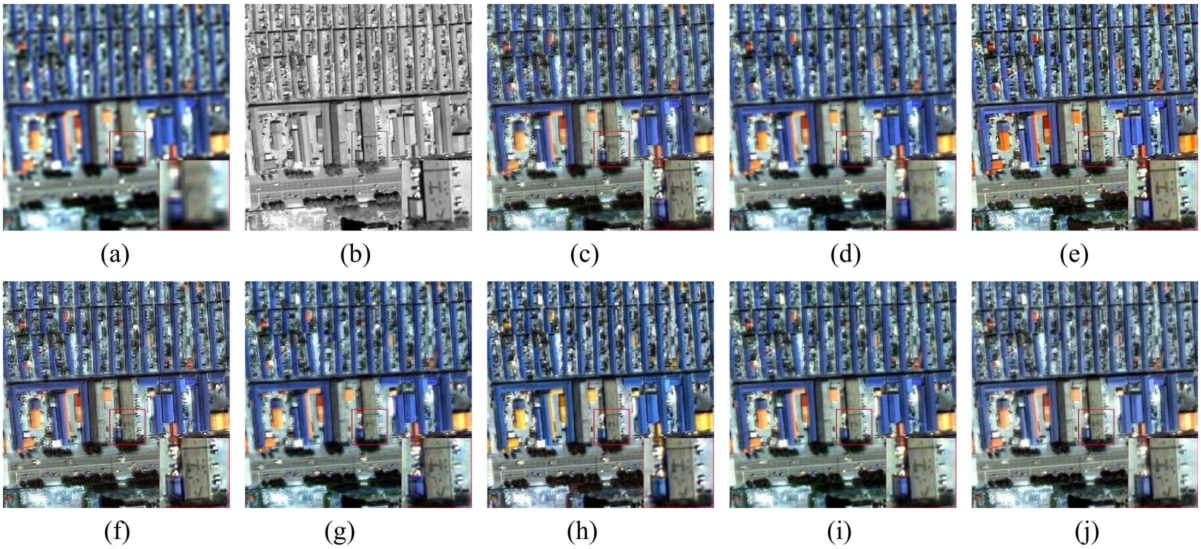

Fig. 8. Fused images by different methods of the second group pleiades data. (a) MS. (b) PAN. (c) AWLP. (d) BDSD_PC. (e) GSA. (f) MTF_GLP. (g) PanNet. (h) PNN. (i) PRACS. (j) Proposed.

TABLE VII
OBJECTIVE EVALUATION OF FUSION RESULTS USING THE SECOND GROUP PLEIADES DATA

	AWLP	BDSD_PC	GS	MTF_GLP	PanNet	PNN	PRACS	Proposed
SSIM	0.6570	0.7264	0.4971	0.5915	**0.7341**	0.6305	0.6121	0.6873
CC	0.8920	0.9353	0.8288	0.8685	0.9032	0.8937	**0.9440**	0.8813
ERGAS	3.8394	2.9837	4.4262	**2.4324**	2.7565	3.8331	4.7257	2.6127
SAM	0.7944	1.8693	1.2838	0.9296	0.9061	1.5945	2.8187	**0.5995**
QNR	0.8048	0.9171	0.8299	0.7569	0.8938	0.8166	0.8777	**0.9337**
D_s	0.0929	0.0533	0.1098	0.1185	0.0672	0.1130	0.0532	**0.0323**
D_λ	0.1127	**0.0315**	0.0678	0.1414	0.0418	0.0794	0.0730	0.0351

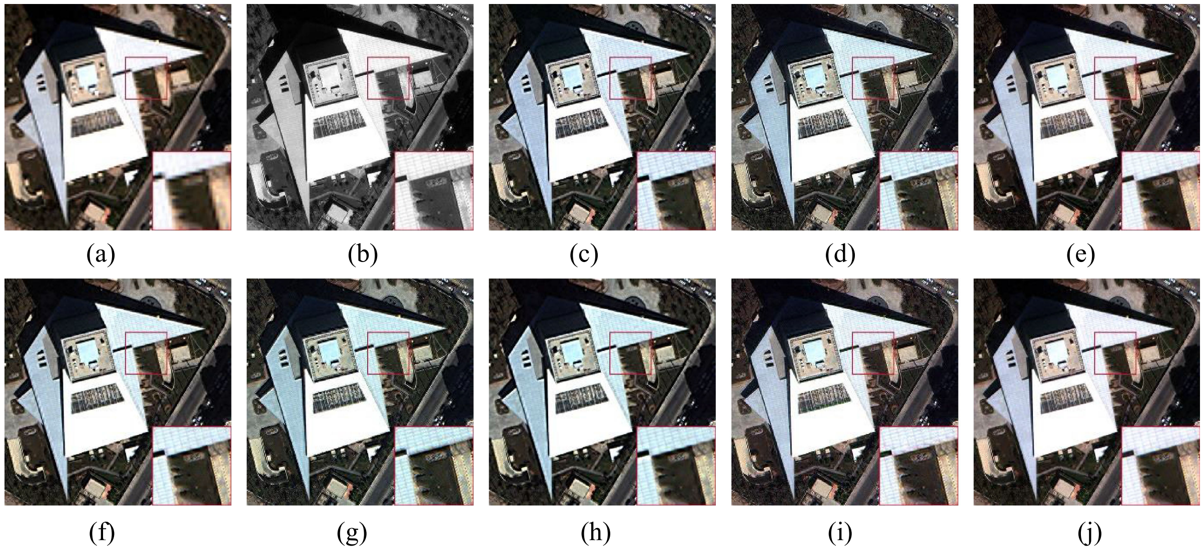

Fig. 9. Fused images by different methods of the third group WorldView-2 data. (a) MS. (b) PAN. (c) AWLP. (d) BDSD_PC. (e) GSA. (f) MTF_GLP. (g) PanNet. (h) PNN. (i) PRACS. (j) Proposed.

TABLE VIII
OBJECTIVE EVALUATION OF FUSION RESULTS USING THE THIRD GROUP WORLDVIEW-2 DATA

	AWLP	BDSD_PC	GS	MTF_GLP	PanNet	PNN	PRACS	Proposed
SSIM	0.8543	0.8676	0.7580	**0.8847**	0.7886	0.8529	0.8473	0.8393
CC	0.9246	0.8743	0.9317	0.9292	0.9341	0.9274	0.9433	**0.9477**
ERGAS	6.4266	8.2625	6.0038	6.3834	5.5097	6.6096	5.5810	**4.6935**
SAM	4.8623	6.1009	4.4342	4.9789	4.5578	4.4706	**4.2950**	4.3043
QNR	0.7863	0.7980	0.8048	0.7610	0.8643	0.7695	0.8125	**0.9305**
D_s	0.1668	0.1608	0.1660	0.1833	0.0663	0.1704	0.1599	**0.0376**
D_λ	0.0563	0.0491	0.0349	0.0682	0.0743	0.0725	**0.0329**	0.0331

2) *Result Analysis for Pleiades Data:* Compared with the MS image by visually in Fig. 8, the roof on the right side of the image of the GS method is darker, it has a serious spectral distortion. Compared with the PAN image, the spatial texture part of the BDSD_PC method is not very clear, which can be seen from the local enlarged image that the capital character H on the ground looks a little fuzzy. The PNN method happens in some spectral distortion. In the middle of the image, the part of the roof is red but it looks like yellow in the fused image of the PNN method. From Table VII, we can see that our proposed method performs best in term of indicators SAM, QNR, and D_s. ERGAS and D_λ rank second which are only a little worse than the optimal ones. The BDSD_PC method performs best in term of indicator D_λ, but its spatial details do not look as rich as the proposed method. The MTF_GLP method performs best in terms of indicators ERGAS and the PanNet performs best in terms of indicators SSIM.

3) *Result Analysis for WorldView-2 Data:* In Fig. 9, from a visual point of view, the fused images of the AWLP, the BDSD_PC, the MTF_GLP, the PanNet, and the PNN methods happen a little spectral distortion. For example, the color of the building looks a little darker than that of the MS. From the local enlarged image, we can see that there are some oblique lines on the edge of the image of the PRACS method. From Table VIII, we can see that our proposed method performs best in term of indicators CC, ERGAS, QNR, and D_s. The SAM and D_λ rank second. The values of SAM and D_λ in the proposed method are 4.3043 and 0.0331, while the best are 4.2950 and 0.0329. The difference is only a little bit. Although the PARCS method performs best in term of indicators SAM and D_λ, it still happens a little spectral distortion in the local enlarged image.

C. Experimental Discuss

Combining the visual fusion results and the quality evaluation indexes, we can consider that the proposed method in this article is better than the most used methods. In the aspect of spatial details, it is well known that the CS-based methods are very effective in spatial enhancement. However, the results of the three groups of experiments proposed by us achieve better spatial enhancement, this is because the spatial distortion index D_s is employed to control the spatial distortion of network output. That is why the proposed method D_s is minimum in the three groups of experiments.

TABLE IX
RUNTIME, PARAMETERS, FLOPS OF THE PROPOSED METHOD

Input type	Runtime	Parameters	FLOPs
WorldView-2	3.75s	161080	1.450G
Pleiades/Gaofen-1	3.24s	139764	1.256G

In the aspect of spectral distortion, both the methods based on MRA and the method based on CNN can keep good spectra to some extent, and the spectral retention ability depends on the fusion framework and parameter settings. At the same time, since MRA based methods usually up-sampling the original MS to the size of PAN, and the CNN-based methods produce simulation training data, which is difficult to make the spectra of the fused image close to the original MS. But in the proposed method, no reference evaluation index D_λ is adopted to control the spectral distortion between the original MS and the fused image. As can be seen from the above-mentioned experiments, the proposed method in this article have an excellent effect on spectral retention.

D. Time Cost and Computational Complexity

All the experiments conducted in this work are performed by using the computer with CPU Corei7/2.20 GHz and 8 GB RAM. The runtime is the average time of predicting 100 images with size 1024×1024. The $64 \times 64 \times n$ image, which has the same size as the training data is adopted to calculate the floating point operations (FLOPs), where the is the number of bands. Table IX shows the runtime, parameters, and FLOPs of our proposed method.

IV. CONCLUSION

Considering the superior performance of the CNN architecture with high learning capacity to form a highly nonlinear transformation, in this article, we improve the SRCNN architecture to get a new pan-sharpening method. In the proposed method, we add a feature enhancement layer which could enhance the obtained features from the feature extraction layer. Moreover, we design a new loss function with no-reference image to simultaneously monitor the spectral loss and spatial loss, which is different from the current loss function used in common in the remote sensing image fusion filed. Thus, the simulated training data in the current deep learning fusion method is not needed. Contrarily, the original MS and PAN are direct as the reference and the training data. Different Gaofen-1, Pleiades, and WorldView-2 remote sensing satellite data with different land covers in different time phases are used to verify the effectiveness of the proposed pan-sharpening method. Seven representative fusion methods and seven evaluation metrics are applied for comparison and evaluation, respectively. The results demonstrate that the proposed method achieves the state-of-art performance in terms of both the visual perception and the objective assessment. The proposed way with no-reference loss function and with no-labeled-fused image pan-sharpening network is probably a new promising starting point in the remote sensing image fusion field.

REFERENCES

[1] T. M. Tu, S. C. Su, H. C. Shyu, and P. S. Huang, "A new look at IHS-like image fusion methods," *Inf. Fusion*, vol. 2, no. 3, pp. 177–186, 2001.

[2] J. Jan, K. S. Veronika, K. Lucie, and M. Jan, "Testing a modified PCA-based sharpening approach for image fusion," *Remote Sens.*, vol. 8, no. 794, pp. 1–25, 2016.

[3] C. Liu, X. Qi, W. Zhang, and X. Huang, "Research of improved Gram-Schmidt image fusion algorithm based on IHS transform," *Eng. Surveying Mapping*, vol. 27, no. 11, pp. 9–14, 2018.

[4] B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS +Pan data," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007.

[5] 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.

[6] 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.

[7] R. Gharbia, A. H. E. Baz, A.E. Hassanien, and M. Tolba, "Remote sensing image fusion approach based on brovey and wavelets transforms," in *Proc. 5th Int. Conf. Innov. Bio-Inspired Comput. Appl. IBICA*, 2014, vol. 303, pp. 311–321.

[8] W. Wang, P. Tang, and C. Zhu, "A wavelet transform based image fusion method," *J. Image Graph.*, vol. 6(A), no. 11, pp. 1130–1136, 2011.

[9] A. A. Suraj, M. Francis, T. S. Kavya, and T. M. Nirmal, "Discrete wavelet transform based image fusion and De-noising in FPGA," *J. Elect. Syst. Inf. Technol.*, vol. 1, pp. 72–81, 2014.

[10] S. Chen, R. Zhang, H. Su, J. Tian, and J. Xia, "SAR and multispectral image fusion using generalized IHS transform based on à trous wavelet and EMD decompositions," *IEEE Sensors J.,* vol. 10, no. 3, pp. 737–745, Mar. 2010.

[11] W. Wang and F. Chang, "A multi-focus image fusion method based on Laplacian pyramid," *J. Comput.*, vol. 6, no. 12, pp. 2559–2566, 2011.

[12] X. Otazu, M. González-Audicana, O. Fors, and J. Murga, "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.

[13] S. Yang, M. Wang, L. Jiao, R. Wu, and Z. Wang, "Image fusion based on a new contourlet packet," *Inf. Fusion*, vol. 11, no. 2, pp. 78–84, 2010.

[14] Z. Zhao and Y. Zheng, "Research of image fusion of multispectral and panchromatic images based on ridgelet transform," *Comput. Eng. Appl.*, vol. 48, no. 15, pp. 164–167, 2012.

[15] W. Kong and J. Liu, "Technique for image fusion based on nonsubsampled shearlet transform and improved pulse-coupled neural network," *Opt. Eng.*, vol. 52, no. 1, 2013, Art. no. 7001.

[16] 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.

[17] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002.

[18] Q. Wei, J. Bioucas-Dias, N. Dobigeon, and J. Tourneret, "Hyperspectral and multispectral image fusion based on a sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 7, pp. 3658–3668, Jul. 2015.

[19] X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 2827–2836, May 2013.

[20] D. F. Hong, N. Yokoya, J. Chanussot, and X. X. Zhu, "An augmented linear mixing model to address spectral variability for hyperspectral unmixing," *IEEE Trans. Image Process.*, vol. 28, no. 4, pp. 1923–1938, Apr. 2019.

[21] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[22] D. Hong *et al.*, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[23] Y. Jing, D. F. Hong, C. Jocelyn, M. Deyu, X. X. Zhu, and Z. B. Xu, "Cross-attention in coupled unmixing nets for unsupervised hyperspectral super-resolution," in *Proc. Euro. Conf. Comput. Vis.*, 2020, pp. 208–224.

[24] L. Gao, D. Hong, J. Yao, B. Zhang, P. Gamba, and J. Chanussot, "Spectral superresolution of multispectral imagery with joint sparse and low-rank learning," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[25] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[26] 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.

[27] Z. Li and C. Cheng, "A CNN-based pan-sharpening method for integrating panchromatic and multispectral images using landsat 8," *Remote Sens.*, vol. 11, no. 22, 2019, Art. no. 2606.

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

[29] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *Proc. Int. Workshop Remote Sens. Intell. Process.*, 2017, pp. 1–4. doi: 10.1109/RSIP.2017.7958807.

[30] 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.

[31] F. Palsson, J. Sveinsson, and M. Ulfarsson, "Sentinel-2 image fusion using a deep residual network," *Remote Sens.*, vol. 10, no. 8, 2018, Art. no. 1290.

[32] Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.

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

[34] A. Azarang and H. Ghassemian, "A new pansharpening method using multi resolution analysis framework and deep neural networks," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal.*, 2017, pp. 1–6.

[35] F. Ozcelik, U. Alganci, E. Sertel, and G. Unal, "Rethinking CNN-based pansharpening: Guided colorization of panchromatic images via GANS," *IEEE Trans. Geosci. Remote Sens.*, to be published.

[36] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.

[37] C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," in *Proc. Euro. Conf. Comput. Vis.*, 2016, vol. 9906, pp. 391–407.

[38] L. Tao, C. Zhu, J. Song, T. Lu, H. Jia, and H. Xie, "Low-light image enhancement using CNN and bright channel prior," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 3215–3219.

[39] 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, 2008.

[40] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2010.

[41] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[42] 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, 1998.

[43] L. Wald, *Data Fusion: Definitions and Architectures–Fusion of Images of Different Spatial Resolutions*. Paris, France: Presses des Mines, 2002, pp. 200.

[44] 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.

[45] G. Vivone *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2014.

Zhangxi Xiong received the B.S. degree in electronic information of science and technology from the Hubei University of Education, Wuhan, China, in 2018. He is currently working toward the M.S. degree in pattern recognition and intelligent systems with Heilongjiang University, Harbin, China. He is currently studying in the Aerospace Information Research Institute, Chinese Academy of Science, Beijing, China.

His current research interests include deep learning, image fusion, and computer vision in remote sensing images.

Qing Guo (Member, IEEE) received the M.Sc. and Ph.D. degrees in optics from the Harbin Institute of Technology, Harbin, China, in 2006 and 2010, respectively.

She joined the Chinese Academy of Sciences, Beijing, China, in 2010. She is currently a Full Professor with Aerospace Information Research Institute, Chinese Academy of Sciences. From 2007 to 2009, she was an exchange Ph.D. student with the Department of Electrical and Computer Engineering, University of Calgary, AB, Canada. From 2014 to 2015, she was a Visiting Scholar with the Institute for Geoinformatics and Remote Sensing, University of Osnabrück, Osnabrück, Germany. Her research interests focus on remote sensing information extraction and processing, including image fusion and deep learning.

Mingliang Liu received the Ph.D. degree in forestry engineering automation from Northeast Forestry University, Harbin, China, in 2017.

He is currently a Full Professor and master's tutor with the School of Electrical Engineering, Heilongjiang University, Harbin, China. His research interests include intelligent detection, fault diagnosis, and signal processing.

An Li received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 1989, and the M.Sc. degree in computer application from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 1992.

He is currently the Director of China Remote Sensing Satellite Ground Station (RSGS), Aerospace Information Research Institute, Chinese Academy of Sciences. He joined RSGS in 1992 and worked on remote sensing data processing. He is currently in charge of the management of the satellite ground segment operation and engineering.