IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received June 23, 2021, accepted July 31, 2021, date of publication August 19, 2021, date of current version August 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104321

Pan-Sharpening Based on Panchromatic Colorization Using WorldView-2

ZHANGXI XIONG[1,2], QING GUO[1], (Member, IEEE), MINGLIANG LIU[2], AND AN LI[1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Harbin 150080, China

Corresponding author: Qing Guo (guoqing@aircas.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61771470, in part by the Key Research Program of Frontier Sciences, Chinese Academy of Sciences, under Grant QYZDY-SSW DQC026, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, under Grant XDA19010401 and Grant XDA19060103, and in part by the Basic Research Fund of Heilongjiang University, Heilongjiang, under Grant kjck201806.

ABSTRACT In order to overcome the lack of the multispectral image (MS) and adequately preserve the spatial information of panchromatic (PAN) image and the spectral information of MS image, this study proposes a method which adds the spectral information of the prior MS to the prior PAN during training, and only the posterior PAN is needed for predicting. Firstly, we introduce the autoencoder model based on image colorization and discuss its feasibility in the field of multi-band remote sensing image pan-sharpening. Then, the image quality evaluation functions including spatial and spectral indexes are formed as the loss function to control the image colorization model. Because the loss function contains spatial and spectral evaluation indexes, it could directly calculate the loss between the network output and the label considering characteristics of remote sensing images. Besides, the training data in our model is original PAN, this means that it is not necessary to make the simulated degraded MS and PAN data for training which is a big difference from most existing deep learning pan-sharpening methods. The new loss function including the spectral and spatial quality instead of the general MSE (mean square error), only the original PAN instead of the simulated degraded MS + PAN to be inputted, only the spectral feature instead of the direct fusion result to be learned, these three aspects change the current learning framework and optimization rule of deep learning pan-sharpening. Finally, thousands of remote sensing images from different scenes are adopted to make the training dataset to verify the effectiveness of the proposed method. In addition, we selected seven representative pan-sharpening algorithms and four widely recognized objective fusion metrics to evaluate and compare the performance on the WorldView-2 experimental data. The results show that the proposed method achieves optimal performance in terms of both the subjective visual effect and the object assessment.

INDEX TERMS Pan-sharpening, deep learning, multispectral image, panchromatic image, image colorization, loss function.

I. INTRODUCTION

Pan-sharpening refers to the fusion of multi-spectral (MS) images with panchromatic (PAN) images to produce high spectral resolution and high spatial resolution images. For the past decades, many pan-sharpening algorithms have been proposed and can be mainly classified into four types: the component substitution (CS), the multi-resolution analysis (MRA), the optimization-based (OB) approaches and the hybrid methods.

Classical pan-sharpening methods including the CS and the MRA methods. The CS methods first rely upon a forward transformation applied to the MS image in order to separate the spatial information with respect to the spectral counterpart. Then, PAN image is adopted to substitute this spatial component. Finally, a spatial enhanced MS can be obtained by the inverse transformation. Some typical examples are the intensity-hue-saturation (IHS) [1], the principal component analysis (PCA) [2], Gram-Schmidt (GS) [3], the Brovey transform (BT) [4], the GS adaptive (GSA) [5] approach and the partial replacement adaptive component substitution (PRACS) [6]. The CS methods are famous for their computational efficiency and robustness to misregistration and

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

aliasing errors. They improve the spatial resolution very well, but most of them usually cause the spectral distortion with varying degrees.

The MRA methods are based on the injection of spatial details that are obtained through a multi-resolution decomposition of the PAN image into the resampled MS bands. Classic MRA methods include the discrete wavelet transform (DWT) [7], the beyond wavelets (e.g., contourlet [8], curvelet [9], shearlet [10]), the Laplacian pyramid (LP) [11], the additive wavelet luminance proportional (AWLP) [12] and the generalized LP based on Gaussian filters matching the modulation transfer function (MTF_GLP) [13]. Compared with the CS methods, the MRA methods generally have superior spectral information preservation but are commonly not satisfactory in terms of the spatial enhancement.

The OB pan-sharpening is posed as an optimization problem between MS and PAN that can be solved by minimizing the loss function with prior constraints. Approaches belonging to this category mainly include the sparse representation (SR) methods and the deep learning methods. SR 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. Li *et al.* learned the dictionaries for PAN and MS and calculated the high spatial resolution MS while retaining the spectral information by integrating the obtained sparse coefficients with the dictionary learning theory [14]. Compared to the MRA methods, the SR methods have super-resolution capability and robustness, and can acquire fused images with less spectral distortion. But it is a challenging problem to find an optimal transformation basis to get the sparsest representation on the transformation basis. Moreover, SR sometimes ignores the intrinsic geometric structure of images.

In addition, a large number of hybrid methods have been proposed, such as IHS + wavelet and IHS + compressed sensing [15] methods. Hybrid methods contain the characteristics of different class methods and achieves good results.

In recent years, the deep learning algorithms have been a hot research field in pan-sharpening. Most of these algorithms learn the mapping relations between low-resolution (LR) multispectral (LRMS) images and high-resolution (HR) multispectral images (HRMS) that are similar to the super resolution convolutional neural network (SRCNN) [16]. Masi *et al.* have proposed a pan-sharpening method by using a three-layer CNN (PNN) [17], which is one of the early applications of convolutional neural networks (CNN) in the pan-sharpening field. After that, a large number of the improved PNN methods [18]–[22] have been proposed, such as adding the residual layer to the CNN and deepening the convolution layer. With the continuous development of the deep learning network, the structure is more complex, and the deeper networks are gradually used to meet the requirements of fusion to maintain higher spectral information and more abundant spatial details. For example, Yang *et al.* have proposed a pan-sharpening method named PanNet that uses the domain specific knowledge to enhance the spectral and structural properties, while training the network in the high-pass domain rather than in the image domain [23]. Ma *et al.* have proposed a pan-sharpening method based on the generative adversarial network (Pan-GAN). In this method, the generator separately establishes the adversarial games with the spectral discriminator and the spatial discriminator, so as to preserve the rich spectral information of multi-spectral images and the spatial information of panchromatic images [24].

The common point of most of the above methods is that they do not fuse the original MS and PAN images directly, but train the simulated degraded data sets made by the Wald protocol [25]. Moreover, the original MS is used as the reference image and the classical loss function such as MSE, mean absolute error (MAE) and cross entropy are adopted to calculate the loss between the reference and the output fusion image. The fused results with small spectral distortion can be obtained by these training methods, but there are still some problems not considered. First of all, they calculate the loss with the network output of simulated training image rather than the original training image, which may ignore some characteristics of the original images. Secondly, the learning relationship between the degraded data and the original MS data may not be the true relationship required for fusion. Thirdly, it wastes a lot of time to make the simulated training data. Therefore, the classical loss function is not optimal for pan-sharpening, and a variety of loss functions have been designed to increase the fitting ability of the model. The common practice is that 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 literatures [26], [27], the weight decaying term and the sparsity term are added to the MSE loss function, respectively. In addition, Choi et al have proposed the spectral-spatial structure (S3) loss function [28] to calculate the spectral loss between the network outputs and the MS targets, and calculate spatial loss between the network outputs and the PAN inputs. Similar to the S3 loss function, Xiong *et al.* have designed a no-reference quality evaluation function based loss function to calculate both the spectral and spatial losses (PLS2, pan-sharpening by using loss function with spatial and spectral quality evaluation function) [29]. But the difference is that Xiong *et al.* have also designed the label making method according to the characteristics of remote sensing images, which can lay in the original PAN and original MS of different spatial resolutions at the same time, and use the real data instead of the simulated data for training.

In view of the above-mentioned methods, there is no method to predict HRMS without MS. Inspired by these successful examples of the designed loss function and the great achievements of image colorization [30], [31], we also consider the pan-sharpening as PAN image colorization. Since there is no ground-truth MS with high spatial resolution as the fusion result reference, we consider using the spatial quality evaluation function and the spectral quality evaluation function as the loss function to control the spatial and spectral quality of the network output at the same time. In this way, our proposed PAN image colorization method uses the original PAN as the network input, and the concatenated up-sampled MS and original PAN as labels. This method not only solves the problem of making simulation data set, but also solves the problem that most deep learning methods do not have HRMS as reference image and cannot get HRMS when MS is missing.

The remainder of this paper is organized as follows. Section 2 describes our proposed framework in detail. Section 3 investigate the optimal parameters and shows both the qualitative and the quantitative analyses through experimental results. Finally, this paper draws conclusions and discusses future work in Section 4.

II. PROPOSED PAN IMAGE COLORIZATION MODEL

A. AUTOENCODER NETWORK

Autoencoder includes two parts: encoder and decoder. The hidden features of the input data can be learned by the encoder part. The new learned features can be reconstructed as close as possible to the original input data by decoder. Autoencoder compresses the input data and extracts the most representative information from the input data. The purpose of autoencoder is to reduce the dimension of input data without losing important features. In short, autoencoder is usually used for dimensionality reduction and feature extraction. Based on this, this paper uses the autoencoder as the feature extractor to extract the features of PAN and generates the new MS image with the same spatial resolution as the PAN.

The autoencoder encodes the input x to get the new features y, and hopes the original input can be reconstructed from the new features. The encoding process is as follows:

$$y = f(wx + b) \tag{1}$$

where w is the weights and b is the bias. Liking the neural network structure, its coding is a linear combination followed by a nonlinear activation function. With the new feature y, the input x can be reconstructed. The decoding process is as follows:

$$x' = g(w'y + b') \tag{2}$$

In order to reconstruct the output as close as possible to the original input, the loss function $L(x, x')$ can be used to train the model.

$$L(x, x') = L(x, g(f(x))) \tag{3}$$

B. AUTOENCODER PAN-SHARPENING MODEL

The purpose of panchromatic colorization is to obtain the MS image with the spatial resolution of PAN. One of the definite characteristics of panchromatic colorization is that the input and the output are the same spatial resolution. In addition, the input and output are different in the color expression, but they have the same internal structure, that is, the same contour, edge and texture. As a result, the structures of the input and the output are almost aligned. Therefore, the architecture of our panchromatic colorization model is also designed around these characteristics.

In the field of image colorization, the creation model always adopts the autoencoder network. In such a network, the encoder will further extract high-level features from the input by passing through each down sampling layer. The decoder mainly samples the extracted features through the encoder layer by layer. Finally, the decoder restore the features to the size of the original image.

In fact, there are a lot of underlying information shared between the input and output. For example, the input and output images share the location information of prominent edges. Therefore, it is advisable to add skip connection in the autoencoder to let the information directly reach the deep network through the shallow network, so as to retain the contour information to the greatest extent. The autoencoder with skip connection is shown in Figure 1.

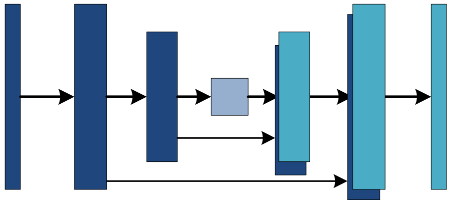

FIGURE 1. The structure of our autoencoder with skip connection.

In this paper, our image creation model based on the autoencoder is shown as Figure 2, in which the autoencoder is used as the image generator to generate the 8 bands MS. In the training stage, the input of the model is a 1 channel PAN image with size 16 × 16. The model mainly includes convolution layer, activation layer and resize-convolution layer. The encoder part is the blue part of Figure.2. The decoder part is the green part of Figure.2. In order to solve the problem that checkerboard artifacts often appears in the dark area of image generated by the deconvolution layer,

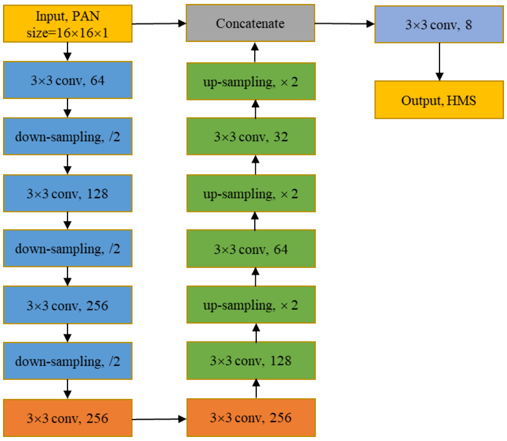

FIGURE 2. The proposed image colorization model.

the resize-convolution layer is used instead of the deconvolution [32]. After convoluting the output of the upper layer with a stride of 1, we up-sample the output of the convolution layer. In addition, the convolution with a stride of 1 and the down-sampling layer are used to replace the convolution with a stride of 2. After decoding, the feature is extracted to the same size as the original PAN image. We supplement the feature information of the lower level by adding skip connections to the higher levels. Finally, we reconstruct the desired HMS image by using 8 filters, where 8 is the band number of MS. The rectified linear unit (ReLU) whose function is max (0, x) is used as the activation function. The detailed architecture and parameters can be found in Figure.2.

C. PAN-SHARPENING ARCHITECTURE

Our proposed architecture consists of two parts: training and predicting. The training part is to learn the super parameters in a supervised manner. The predicting part is to generate the HMS through the learned super parameters during the training stage.

Figure 3 shows the workflow of training and predicting of our proposed panchromatic image colorization model. In the training stage, the original PAN is the input of model, and the output is a n-bands image which corresponds to the number of MS bands. Then, losses are calculated between the output and the label. It is worth mentioning that the label is n + 1 bands image which consists of the n-bands up-sampled MS and the original PAN. During the training stage, the stochastic gradient descent algorithm (SGD) and the back propagation (BP) are utilized to iteratively learn all of the parameters (w, b) in the network for optimal allocation.

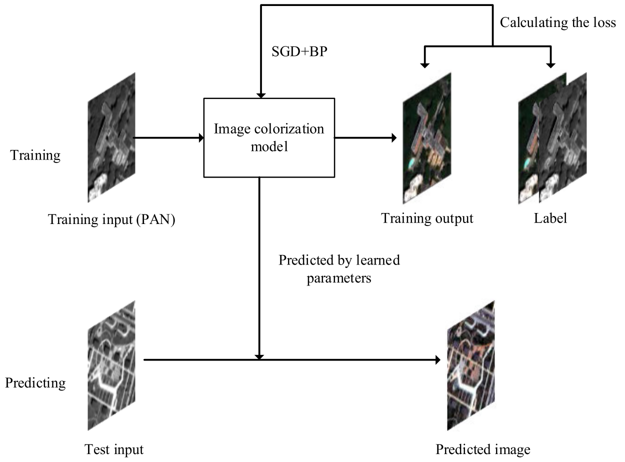

FIGURE 3. The proposed pan-sharpening architecture.

The image quality evaluation functions—the spectral angle mapper (SAM) [33] and the universal image quality index (UQI) [34] are adopted as the loss function.

SAM is defined as:

$$SAM(v, \hat{v}) = \arccos\left(\frac{\langle v, \hat{v}\rangle}{\|v\|_2 * \|\hat{v}\|_2}\right) \tag{4}$$

SAM denotes the absolute value of the spectral angle between two vectors. In the formula (4), 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.

UQI is defined as:

$$UQI \triangleq \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \times \frac{2 \cdot \bar{x} \cdot \bar{y}}{\bar{x}^2 + \bar{y}^2} \times \frac{2 \cdot \sigma_x \cdot \sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{5}$$

where σ_{xy} denotes the covariance between x and y, $\bar{x}$ and $\bar{y}$ are the means, σ_x^2 and σ_y^2 are the variance of x and y, respectively. The first one is the correlation coefficient (CC) between x and y, which ranges in [-1, 1]. According to Cauchy-Schwartz inequality, the second one and the third one range in [0, 1]. Hence, the dynamic range of UQI is [-1, 1], and the best value $UQI = 1$ is achieved when $x = y$ for all pixels. UQI is the basis of the calculation of the spatial quality evaluation function D_s and spectral quality evaluation function D_λ.

When calculate the losses, the n + 1 bands label is first divided into two tensors: the n-bands up-sampled MS named UMS, and the other one band original PAN. Then, $\tilde{F}$ is utilized to represent the model output. Therefore, the loss function $L(w, b)$ is defined as:

$$sam = SAM(UMS, \tilde{F}) \tag{6}$$

$$Q = \frac{1}{n}\sum_{i=1}^{n} UQI(\tilde{F}_i, PAN) \tag{7}$$

$$L(w, b) = \alpha \cdot sam + \beta \cdot (-Q) \tag{8}$$

where (w, b) is the set of all involved super parameters of the proposed pan-sharpening architecture, which are named filter weights and biases. α and β are the weights added to sam and $-Q$, respectively. Since the best values of Q and sam are the maximum and the minimum of the value range, respectively, in order to make its value direction tend to be the same, we multiply Q by minus one.

D. IMPLEMENT DETAILS

The algorithm in this paper is implemented using Tensorflow. We crop the training data set to many samples of the same size as 16 × 16. For the setting of experimental parameters, we did a quantitative analysis experiment, and used the experimental results to determine the best parameters within a preselected range. According to the experimental results, we set the convolution kernel size to 3 × 3, the training data size is set to 16 × 16, α and β are set $\alpha = 0.7$ and $\beta = 0.3$, respectively. In addition, according to experience, the SGD optimizer is used to reduce the randomness, the momentum is set to 0.9, and the learning rate is set to 10^{-4}. The number of iterations is fixed to 4×10^3. The batch size is set to 32. The training process of our network costs 10 h roughly.

In the predicting stage, we choose PAN that is not in the training data from the same source satellite image as the test input. Finally, the fused HMS image is predicted through the learned super parameters during the training stage.

III. EXPERIMENTATION AND ANALYSIS

A. DATASETS

In this paper, the WorldView-2 satellite images are used in experiments. WorldView-2 satellite operates at an altitude of 770 km and an orbit inclination of 98 degrees. Its orbit type is the solar synchronous orbit with a repetition period of 93.4 minutes. WorldView-2 satellite provides 1 band PAN and 8 bands MS. Table 1 shows the band information, where the wavelength range (in nm) and the resolution (m/pixel) are reported. We train the networks using 40,000 PAN/label patch pairs of size 16 × 16 from datasets of WorldView-2 satellite.

TABLE 1. The band information of WorldView-2 satellite image.

Spectral Band	Wavelength	Resolution
Coastal	400-450	1.84
Blue	450-510	1.84
Green	510-580	1.84
Yellow	585-625	1.84
Red	630-690	1.84
Red Edge	705-745	1.84
Near Infrared 1	770-895	1.84
Near Infrared 2	860-1040	1.84
Panchromatic	450-800	0.46

B. METHODS FOR COMPARISON AND OBJECTIVE EVALUATION METRICS

In this paper, seven representative pan-sharpening algorithms are selected for comparison, which are GSA [5], PRACS [6], MTF_GLP [13], PNN [17], PanNet [23], TACNN [22] and PLS2 [29]. Considering that the comparison method should be comprehensive, therefore, in these comparison methods, the GSA and the PRACS methods belong to the category of CS, the MTF_GLP method belong to the category of MRA, and the PNN, PanNet, TACNN and PLS2 methods belong to the category of OB. For the objective quantitative evaluation in the experiments, four widely recognized objective fusion metrics are adopted, which are erreur relative globale adimensionnelle de synthèse (ERGAS) [35], spectral distortion index D_λ [36], [37], structural similarity index (SSIM) [38] and spatial distortion index D_s [36], [37]. More details are given as follows:

ERGAS expresses the error between the up-sampled MS and the fused image, which is defined as:

$$ERGAS = 100 \cdot \frac{h}{l}\sqrt{\frac{1}{k}\sum_{i=1}^{k}\left[\frac{RMSE(i)}{\mu(i)}\right]^2} \tag{9}$$

$$RMSE(F, MS) = \sqrt{\frac{1}{M*N}\sum_{u=1}^{M}\sum_{v=1}^{N}[F(u,v) - MS(u,v)]^2} \tag{10}$$

where h and l are the spatial resolutions of the PAN and MS images, respectively. k is the number of bands of the fused image. $RMSE$ is the root mean squared error. RMSE gives the standard measure of difference between F and MS. $\mu(i)$ denotes the mean of i_{th} band of reference MS image. The smaller the ERGAS is, the closer the fused image is to the up-sampled MS. The best ERGAS value is 0.

D_λ is used to measure the degree of spectral distortion, and D_s is used to measure the degree of spatial distortion. They are defined as (11) and (12), as shown at the bottom of the next page, where UQI is defined in the above formula (5). $UQI(fused_l, fused_r)$ is the quality index values between the l_{th} band and the r_{th} band of fused image, $UQI(MS_l, MS_r)$ calculates the quality index values between the l_{th} band and the r_{th} band of MS image, $UQI(fused_l, P)$ is the quality index values between the fused image l_{th} band and PAN, $UQI(MS_l, \tilde{P})$ is the quality index values between the MS image l_{th} band and the degraded PAN image. L is the number of MS bands. p and q are typically set to 1. 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.

SSIM reflects the structural similarity of two image. It is defined as:

$$SSIM(F, MS) = \frac{(2 \cdot \mu_F \cdot \mu_{MS} + c_1) * (2 \cdot \sigma_{FMS} + c_2)}{(\mu_F^2 + \mu_{MS}^2 + c_1) * (\sigma_F^2 + \sigma_{MS}^2 + c_2)} \tag{13}$$

where μ_F and μ_{MS} denote the mean values of the fused image F and the MS image, respectively. σ_F^2 and σ_{MS}^2 represent the variances of the fused image F and the MS image, respectively. The covariance of images is represented by σ_{FMS}, and c is a constant. The bigger the $SSIM$, the more similarities between the images. The best value of $SSIM$ is 1.

C. NETWORK PARAMETER SETTING

In the training stage, there are some parameters affecting the performance of our proposed architecture output. Some primary parameters will be learned. The other parameters, such as the convolution kernel size and the input training data size, are set unchanged when one parameter is evaluated and investigated. The details of selecting the optimal parameters are as follows:

1) SELECTION OF CONVOLUTION KERNEL SIZE

As mentioned in Section 2.2, in the decoding part, we use the resize-convolution layer to replace the deconvolution layer with the stride of 2 to avoid checkerboard artifacts. In the selection of the convolution kernel size, we choose the odd number for the convolution kernel size, because the odd number has center point which is convenient for padding. Moreover, compared with the even number of the kernel size, the odd kernel size is more sensitive to edges and lines, that can extract the edge information more effectively. Therefore, it is set to 3 × 3, 5 × 5, . . . , 11 × 11 in our implementation. The effect of the convolution kernel size in SSIM, ERGAS, D_s and D_λ is summarized in Figure 4. From Figure 4, we can find that with the increasing of the kernel size, the four evaluation indexes have no

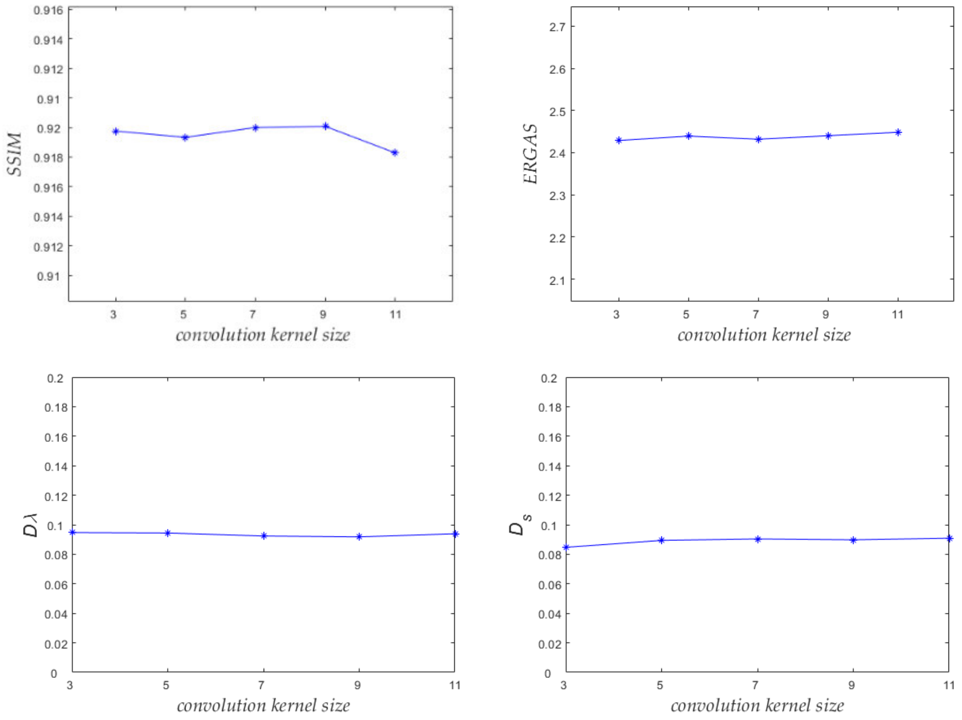

FIGURE 4. **Quality evaluation of the fused image with different convolution kernel sizes.**

obvious change. In particular, ERGAS and D_λ are almost stable at 2.33 and 0.09, respectively, while SSIM and D_s have little fluctuation range. This shows that the convolution kernel size has little effect on our model, but the computational complexity will increase with the increasing of kernel size. Therefore, in order to maintain the high computing performance of our model, the convolution kernel size should be as small as possible. As shown in Figure 4, we set the convolution kernel size to 3 × 3.

2) SELECTION OF TRAINING DATA SIZE

Our model coding part contains three down-sampling layers, which means that the size of the encoding result is one eighth of the input image. In order to make it meaningful, the size of the model training data must be a positive integer multiple of 8. To evaluate the effect of the training data size on our proposed model, it is set in our implementation to 8 × 8, 16 × 16, 24 × 24 , . . . , 64 × 64. The effect of the convolution kernel size in SSIM, ERGAS, D_s and D_λ is summarized in Figure 5. As can be seen from Figure 5, when the size of the training data increases from 8 × 8 to 16 × 16, the four evaluation indexes all change obviously, among which ERGAS, D_s and D_λ are decreased and SSIM is increased. However, when the training data size increases from 16 × 16 to 64 × 64, ERGAS and D_s show an upward trend with small fluctuation range, D_λ changes not obviously, SSIM begins to fluctuate but remains stable around 0.92. Overall, when the input image size is 8 × 8, it will result in the incomplete feature extraction. When the training data size is increased to 16 × 16, the quality evaluation result is obviously improved. When the training data size is increased again, the quality evaluation result will not be significantly improved. Therefore, in this experiment, the training data size is set to 16 × 16, which also can reduce the computational complexity of the model.

3) ROLE OF α AND β IN LOSS FUNCTION

During the training stages, the setting of α and β is the key in our research. When α is set to 1, the loss function is *sam*

$$D_\lambda \triangleq \sqrt[p]{\frac{1}{L(L-1)}\sum_{l=1}^{L}\sum_{\underbrace{r=1}_{r\neq l}}^{L}\left|UQI(fused_l, fused_r) - UQI(MS_l, MS_r)\right|^p} \tag{11}$$

$$D_s \triangleq \sqrt[q]{\frac{1}{L}\sum_{l=1}^{L}\left|UQI(fused_l, P) - UQI(MS_l, \tilde{P})\right|^q} \tag{12}$$

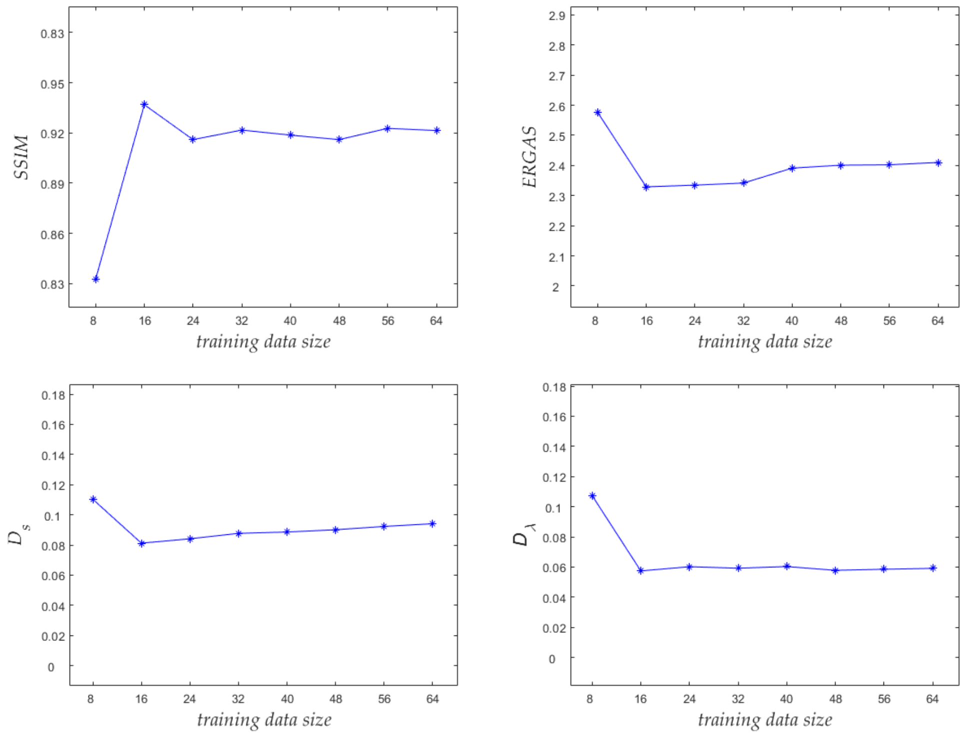

FIGURE 5. Quality evaluation of the fused image with different training data sizes.

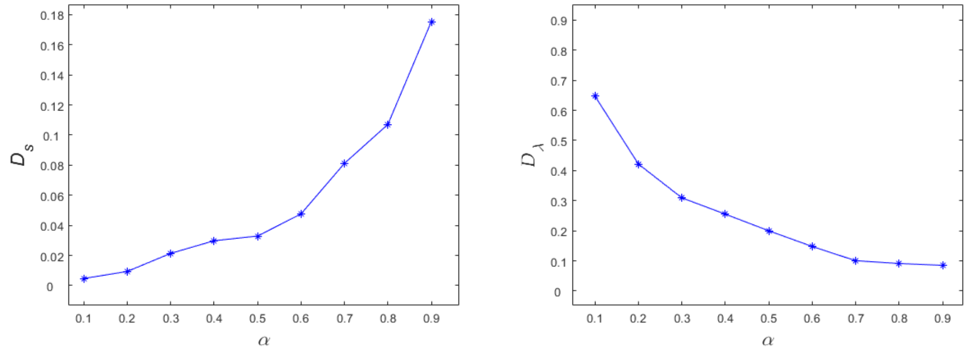

FIGURE 6. Quality evaluation of fused image with different weight.

which will cause the network output to be infinitely close to the up-sampled MS. When α is set to 0, the loss function is $-Q$ which will cause each band of the network output to be infinitely close to the PAN. In order to make the network output keep both the spectral characteristics of MS and the spatial resolution of PAN as much as possible, it is necessary to balance the values of α and β. Therefore, a comparative experiment is used to determine the optimal values of α and β in our loss function. In the experiment, α is set to 0.9, 0.8, ..., 0.1, the corresponding β is set to 0.1, 0.2, ..., 0.9. The spatial distortion index D_s and the spectral distortion index D_λ are adopted to measure the network output to determine the optimal values of α and β. Figure 6 shows the performance fluctuation in D_s and D_λ. In order to observe the influence of different α on the experimental results, Figure 7 displays the part of network output. From Figure 6, we can find that with the increasing of α, the value of D_λ tends to decrease, but when α is greater than or equal to 0.7, the decreasing trend of D_λ slows obviously down, and the value of D_λ is less than 0.1. At the same time, D_s increases with the increasing of α, and the increasing trend increases obviously when α is greater than 0.6, even when α is equal to 0.7, D_s does not exceed 0.1. This shows that the loss function Q has an excellent ability to control the spatial loss, even if Q is given a small weight. Combined with the effect in Figure 7, the weight α of *sam* in this experiment is taken as 0.7, and the corresponding β is taken as 0.3.

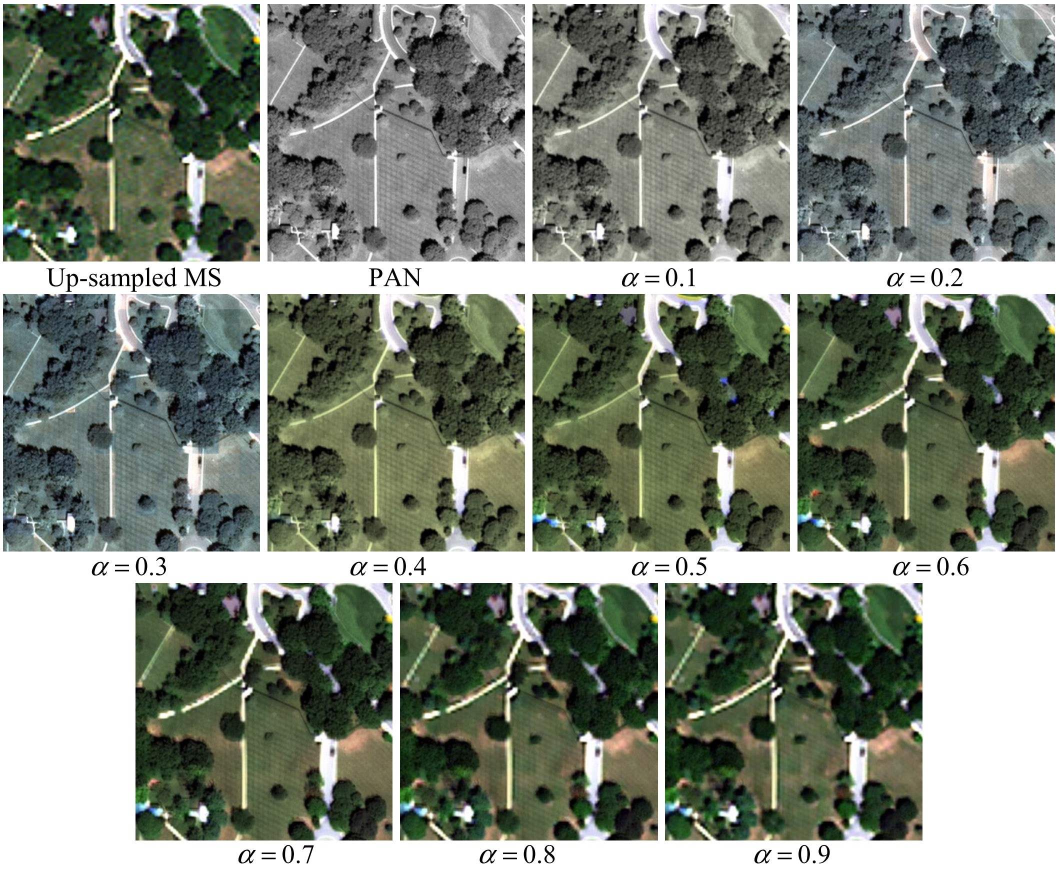

FIGURE 7. The network output with different weights.

TABLE 2. The objective evaluation of fused images.

Methods	ERGAS	SSIM	D_s	D_λ
GSA	3.8778	0.9108	0.1217	0.1728
PRACS	2.4492	0.9334	0.1509	0.1292
MTF_GLP	2.7275	0.9176	0.1264	0.1479
PNN	2.5594	0.9226	0.1433	0.1643
PanNet	2.6078	0.9053	0.1172	0.1689
TACNN	2.4832	<u>0.9336</u>	0.1131	0.1051
PLS2	<u>2.3671</u>	0.9228	<u>0.1104</u>	<u>0.1033</u>
Proposed	**2.3029**	**0.9347**	**0.1022**	**0.0816**

TABLE 3. List of the main acronyms.

Acronyms	Description
MS	multispectral image
PAN	panchromatic
MSE	mean square error
CS	component substitution
MRA	multi-resolution analysis
OB	optimization-based
IHS	intensity-hue-saturation
PCA	principal component analysis
GS	Gram-Schmidt
BT	Brovey transform
PRACS	partial replacement adaptive component substitution
DWT	discrete wavelet transform
LP	Laplacian pyramid
AWLP	additive wavelet luminance proportional
SR	sparse representation
LR	low-resolution
HR	high-resolution
CNN	convolutional neural networks
SRCNN	super resolution convolutional neural network
MAE	mean absolute error
S3	spectral-spatial structure
ReLU	rectified linear unit
SGD	stochastic gradient descent
BP	back propagation
SAM	spectral angle mapper
UQI	universal image quality index
ERGAS	erreur relative globale adimensionnelle de synthèse
SSIM	structural similarity index

D. EXPERIMENTAL RESULTS AND ANALYSIS

In order to prove the spectral retention performance of our proposed model, two groups band combination of the fused image are displayed. One group is natural color composition including the red, green and blue bands, and the other is pseudo color synthesis including the near infrared 2, near infrared 1 and red edge bands. The size of the displayed image is 400 × 400.

For the quantitative analysis, Table 2 shows the average objective evaluation values. The best values are marked in bold font and the second best is underlined. In Table 2, the proposed PAN image colorization method outperforms other comparison methods with values in ERGAS, SSIM, D_s and D_λ of 2.3029, 0.9347, 0.1022 and 0.0816 on average, respectively. The PLS2 method performs the second best ERGAS (2.3671), D_s (0.1104) and D_λ (0.1033). The TACNN method performs the second best SSIM which is 0.9336.

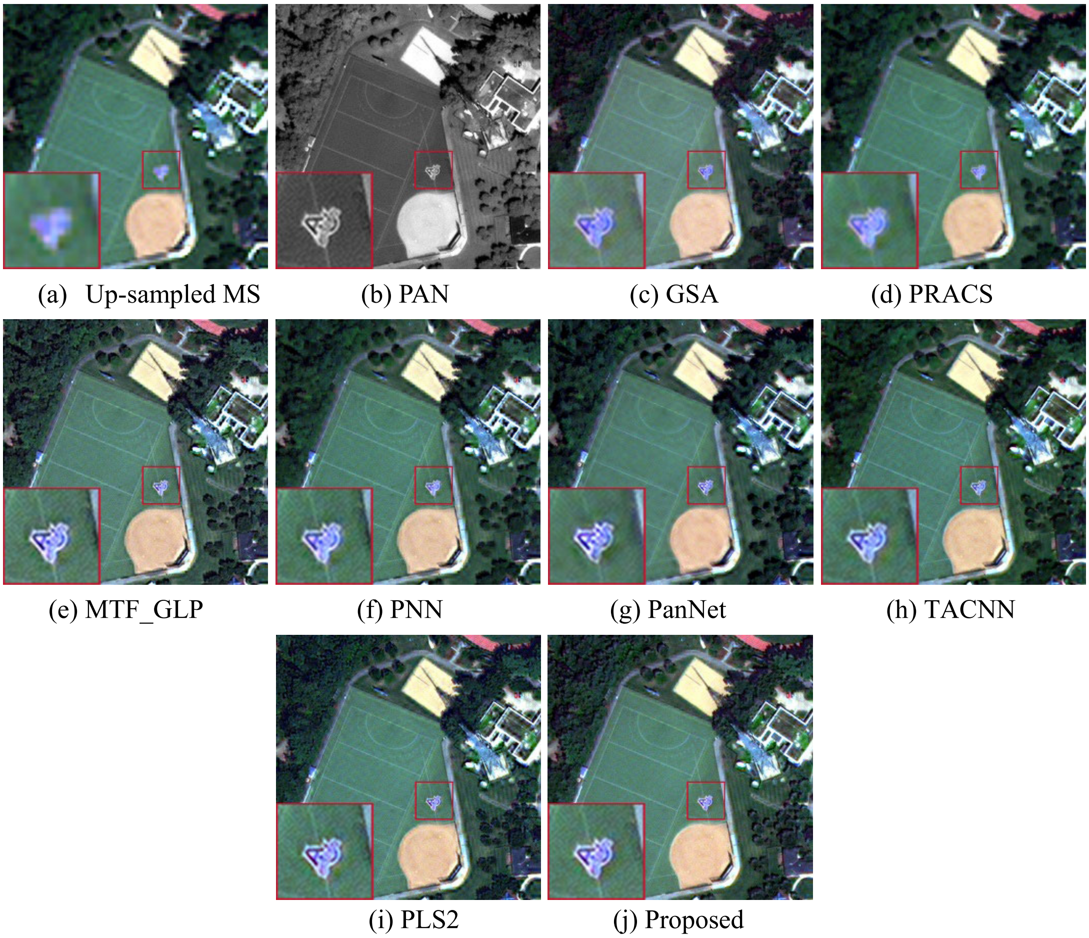

FIGURE 8. The fused images by different methods in the natural color composition.

It can be seen from Table 2 that PLS2 method and the proposed method perform better than other methods. The reason may be that these two methods design loss function according to the characteristics of remote sensing images to control the spatial and spectral losses of the network output at the same time. Compared with PLS2 method, our loss function calculation is simpler, which means less distortion points will be produced in complex calculation. In addition, a group of comparative experiments are conducted to determine the optimal weight of the loss function, while PLS2 method is only a simple comparison, and this may be the reason that the proposed method performs a little better than the PLS2 method. The GSA method has the worst performance in the spectral quantitative evaluation metric ERGAS and D_λ, probably because it belongs to the category of CS, and some information may be lost in the process of component substitution. The PRACS and MTF_GLP methods are close to the PNN and PanNet methods in some quantitative evaluation metrics. From this above, we could find that although the neural network has strong fitting ability, in order to give full play to the potential of neural network, we need to have a deeper understanding of the problem itself and be sensitive enough to the characteristics of remote sensing images so as to design a network structure suitable for this problem.

Figure. 8 and Figure. 9 show the visual comparisons among different methods with the natural color composition and the pseudo color synthesis, respectively. In Figure. 8, from a visual point of view, the fused images of GSA (Figure. 8(c)) method happens a very serious spectral distortion. In Figure 8(c), the trees in the up-left corner of the image look darker than that in the up-sampled MS. The spectral information of Figure 8(d) is maintained well, but compared with PAN, the local enlarged image of PRACS result looks a little blurred, and the spatial detail is not rich. In the local enlarged images of MTF_GLP, PanNet and PNN results, the capital character A is surrounded by some black areas, which show the three methods have partial spectral distortion. The results of the TACNN method, PLS2 method and the proposed method seem to improve the spatial resolution while maintaining the spectral information of MS. In Figure 9, compared with the up-sampled MS, it is obvious that the

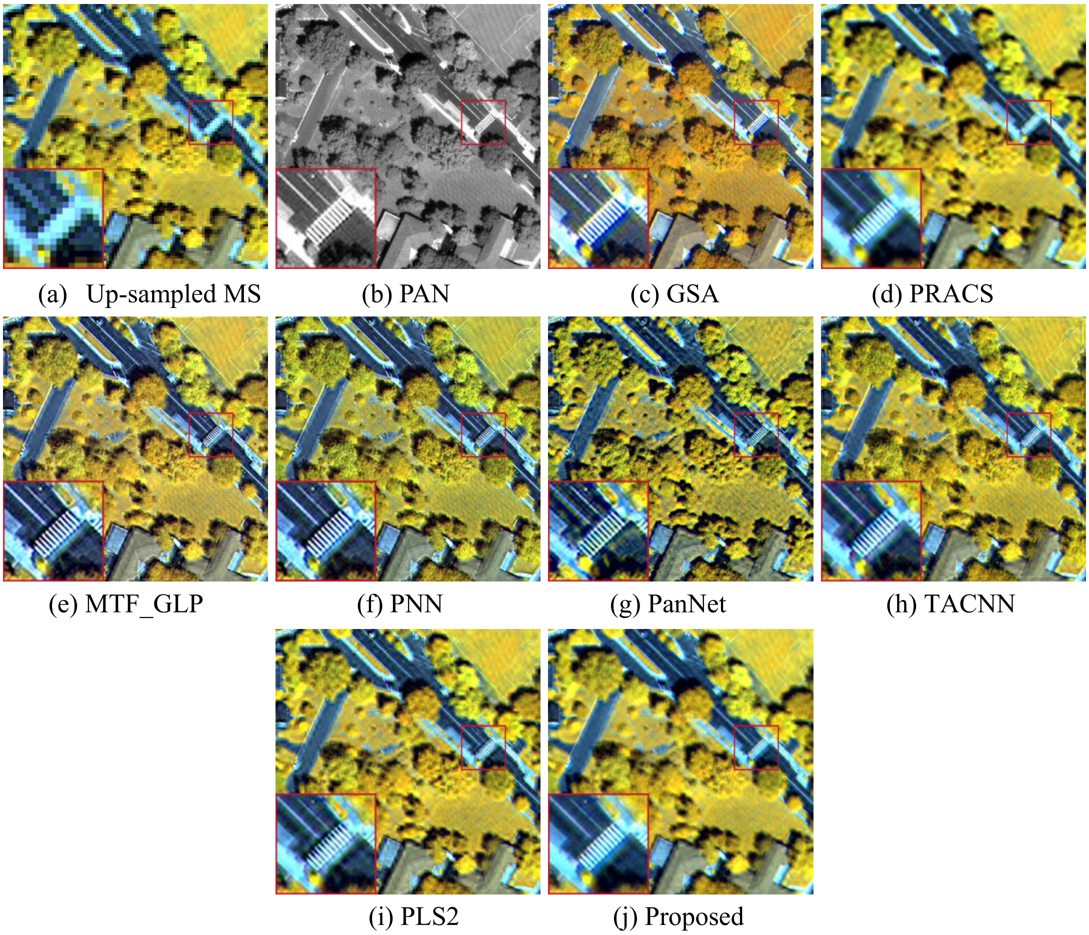

FIGURE 9. The fused images by different methods in the pseudo color synthesis.

fused images of GSA and MTF_GLP methods look a bit more yellow. For the PNN result, the trees in the middle do not look as bright as the up-sampled MS. Although the spatial resolution of PanNet results has been greatly improved, it seems the fused images are added random noise. The PRACS and the proposed methods maintain good spectra quality, but in the aspect of the spatial detail improvement, the fused image of PRACS method is not as good as the proposed method.

IV. CONCLUSION

In this paper, a novel pan-sharpening structure which is a variation of the normal gray image colorization model is proposed. The proposed idea learns the spatial and spectral feature of fusion, not the direct fusion result, while based on the original image to be fused, not the simulated degraded image. Compared with the traditional gray image colorization, the biggest difference is that both the spectral and the spatial quality evaluation functions are simultaneously introduced as the loss function, which changes the learning target and the learning framework based on the original data. Specifically, SAM and UQI are adopted to calculate the spectral and spatial loss, respectively. This makes the network output fusion result has high spectral similarity with MS and high spatial similarity with PAN. In addition, we discuss and test the influence of different weights and model parameters on the fusion results. By using the proposed model with the designed loss function, the production of the training data and labels becomes simple, and there is no need to make the simulated degraded MS+PAN data, which is the biggest difference from the most deep learning remote sensing image fusion algorithms. After the training, the high spatial resolution MS can be obtained only by inputting PAN. This means that pan-sharpening can also be done when MS is missing. Seven representative fusion methods and four 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 panchromatic image colorization model with the designed loss function is probably a new promising starting point in the remote sensing image fusion field.

REFERENCES

[1] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman, ''An adaptive IHS pan-sharpening method,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746–750, Oct. 2010, doi: 10.1109/LGRS.2010.2046715.

[2] V. P. Shah, N. H. Younan, and R. L. King, ''An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008, doi: 10.1109/TGRS.2008.916211.

[3] M. D. Mura, G. Vivone, R. Restaino, P. Addesso, and J. Chanussot, ''Global and local gram-Schmidt methods for hyperspectral pansharpening,'' in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Milan, Italy, Jul. 2015, pp. 37–40, doi: 10.1109/IGARSS.2015.7325691.

[4] R. Gharbia, A. H. E. Baz, A. E. Hassanien, and M. Tolba, ''Remote sensing image fusion approach based on Brovey and wavelets transforms,'' *Proc. 5th Int. Conf. Innov. Bio-Inspired Comput. Appl.*, vol. 303, 2014, pp. 311–321, doi: 10.1007/978-3-319-08156-4_31.

[5] B. Aiazzi, S. Baronti, and M. Selva, ''Improving component substitution pansharpening through multivariate regression of MS +Pan data,'' in *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3230–3239, Oct. 2007, doi: 10.1109/TGRS.2007.901007.

[6] J. Choi, K. Yu, and Y. Kim, ''A new adaptive component-substitution-based satellite image fusion by using partial replacement,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011, doi: 10.1109/TGRS.2010.2051674.

[7] Y. Yang, C.-Z. Han, and D.-Q. Han, ''A structure information based image fusion algorithm using IHS and discrete wavelet transform,'' in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, Beijing, China, Nov. 2007, pp. 859–864, doi: 10.1109/ICWAPR.2007.4420720.

[8] V. P. Shah, N. H. Younan, and R. King, ''Pan-sharpening via the contourlet transform,'' in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Barcelona, Spain, Jul. 2007, pp. 310–313, doi: 10.1109/IGARSS.2007.4422792.

[9] S. Ren, J. Cheng, and M. Li, ''Multiresolution fusion of pan and MS images based on the curvelet transform,'' in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Honolulu, HI, USA, Jul. 2010, pp. 472–475, doi: 10.1109/IGARSS.2010.5652557.

[10] 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, 2019, doi: 10.1109/ACCESS.2019.2910656.

[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, doi: 10.4304/jcp.6.12.2559-2566.

[12] X. Otazu, M. González-Audícana, O. Fors, and J. Núñez, ''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, doi: 10.1109/TGRS.2005.856106.

[13] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, ''Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Oct. 2002, doi: 10.1109/TGRS.2002.803623.

[14] S. Li, H. Yin, and L. Fang, ''Remote sensing image fusion via sparse representations over learned dictionaries,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779–4789, Sep. 2013, doi: 10.1109/TGRS.2012.2230332.

[15] C. Yang, Q. Zhan, H. Liu, and R. Ma, ''An IHS-based pan-sharpening method for spectral fidelity improvement using ripplet transform and compressed sensing,'' *Sensors*, vol. 18, no. 11, p. 3624, Oct. 2018, doi: 10.3390/s18113624.

[16] C. Dong, C. C. Loy, and X. Tang, ''Accelerating the super-resolution convolutional neural network,'' in *Proc. ECCV*, vol. 9906, 2016, pp. 391–407, doi: 10.1007/978-3-319-46475-6_25.

[17] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, ''Pansharpening by convolutional neural networks,'' *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016, doi: 10.3390/rs8070594.

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

[19] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, ''Boosting the accuracy of multispectral image pansharpening by learning a deep residual network,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017, doi: 10.1109/LGRS.2017.2736020.

[20] 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, p. 2606, Nov. 2019, doi: 10.3390/rs11222606.

[21] T. Wang, L. Yang, and L. Xu, ''Multispectral images pan-sharpening based on atrous convolution network and deep residual network,'' in *Proc. 3rd Int. Conf. Adv. Image Process.*, Nov. 2019, p. 71–75, doi: 10.1145/3373419.3373461.

[22] 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, doi: 10.1109/TGRS.2018.2817393.

[23] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, ''PanNet: A deep network architecture for pan-sharpening,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 1753–1761, doi: 10.1109/ICCV.2017.193.

[24] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, ''Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion,'' *Inf. Fusion*, vol. 62, pp. 110–120, Oct. 2020, doi: 10.1016/j.inffus.2020.04.006.

[25] 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, doi: 10.1016/S0924-2716(97)00008-7.

[26] 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, doi: 10.1109/LGRS.2014.2376034.

[27] 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. (IPRIA)*, Shahrekord, Iran, Apr. 2017, pp. 1–6, doi: 10.1109/PRIA.2017.7983017.

[28] J.-S. Choi, Y. Kim, and M. Kim, ''S3: A spectral-spatial structure loss for pan-sharpening networks,'' *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 5, pp. 829–833, May 2020, doi: 10.1109/LGRS.2019.2934493.

[29] Z. Xiong, Q. Guo, M. Liu, and A. Li, ''Pan-sharpening based on convolutional neural network by using the loss function with no-reference,'' *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 897–906, 2021, doi: 10.1109/JSTARS.2020.3038057.

[30] R. Zhang, P. Isola, and A. A. Efros, ''Colorful image colorization,'' in *Proc. ECCV*, 2016, p. 9907, doi: 10.1007/978-3-319-46487-9_40.

[31] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, ''Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5801–5810, doi 10.1109/CVPR42600.2020.00584.

[32] A. Odena, V. Dumoulin, and C. Olah, ''Deconvolution and checkerboard artifacts,'' *Distill*, vol. 1, no. 10, p. e3, Oct. 2016, doi: 10.23915/distill.00003.

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

[34] Z. Wang and A. C. Bovik, ''A universal image quality index,'' *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002, doi: 10.1109/97.995823.

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

[36] G. Vivone, L. Alparone, J. Chanussot, M. D. Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, ''A critical comparison among pansharpening algorithms,'' *IEEE Trans. Geosci. Remote Sensing.*, vol. 53, no. 5, pp. 2565–2586, Dec. 2014, doi: 10.1109/TGRS.2014.2361734.

[37] G. Vivone, M. Dalla Mura, A. Garzelli, R. Restaino, G. Scarpa, M. O. Ulfarsson, L. Alparone, and J. Chanussot, ''A new benchmark based on recent advances in multispectral pansharpening: Revisiting pansharpening with classical and emerging pansharpening methods,'' *IEEE Geosci. Remote Sens. Mag.*, vol. 9, no. 1, pp. 53–81, Mar. 2021, doi: 10.1109/MGRS.2020.3019315.

[38] 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, doi: 10.1109/TIP.2003.819861.

ZHANGXI XIONG is currently pursuing the M.Sc. degree in pattern recognition and intelligent systems with Heilongjiang University. He is studying at the Aerospace Information Research Institute, Chinese Academy of Sciences. 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 Harbin Institute of Technology, Harbin, China, in 2006 and 2010, respectively. From 2007 to 2009, she was an Exchange Ph.D. Student with the Department of Electrical and Computer Engineering, University of Calgary, AB, Canada. In 2010, she joined the Chinese Academy of Sciences, where she is currently a Full Professor with the Aerospace Information Research Institute. From 2014 to 2015, she was a Visiting Scholar with the Institute for Geoinformatics and Remote Sensing, University of Osnabrück, Germany. Her research interests include remote sensing information extraction and processing, including the image fusion and deep learning.

MINGLIANG LIU received the Ph.D. degree in forestry engineering automation from Northeast Forestry University, in 2017. He is currently a Full Professor and a Master's Tutor with the School of Electrical Engineering, Heilongjiang University. 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, in 1992. He is currently a Full Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include remote sensing data processing and satellite ground system management.