



Remote Sensing Image Fusion Using Multi-Scale Convolutional Neural Network

Wei Shi¹ · ChaoBen Du² · BingBing Gao² · JiNing Yan³

Received: 13 November 2019 / Accepted: 10 March 2021 / Published online: 23 March 2021
© Indian Society of Remote Sensing 2021

Abstract

In this paper, a novel remote sensing (RS) image fusion algorithm based on Multi-scale convolutional neural network is proposed. The most important innovation is that the proposed remote sensing image fusion method utilizes a set of convolutional neural networks (CNN) to perform multi-scale image analysis on each band of a multispectral image in order to extract the typical characteristics of different band of multispectral images. In addition, to prevent losing the information of the original image, the max-pooling layer of the traditional CNN is replaced with a standard convolutional layer, and the standard convolutional layer has one step size of 2. The RS image fusion results presented in this paper demonstrate that the proposed method is not only competitive with the most advanced methods, but also superior to other classical methods.

Keywords Remote-sensing · Convolutional neural network · Image fusion

Introduction

Electromagnetic waves reflected on the Earth's surface can be detected and recorded by satellite sensor, while remote sensing (RS) images are the carriers for storing such information (Deng et al., 2019). The purpose of RS image is to extract information on the structure and content of the Earth's surface, to monitor the environment and climate (Plowright et al., 2017; Posselt et al. 2012) and to detect and classify land cover changes (Matikainen et al., 2017; Zhang et al., 2016). Mainstream Earth observation satellites (e.g., Ikonos, Landsat, Gaofen-1 and Quick-Bird) can simultaneously capture panchromatic (PAN) images and multi-spectral (MS) images in the same coverage area. PAN images contain rich spatial resolution information that helps distinguish different substances in the scene. Due to its strong recognition ability, it has been widely used,

especially for Earth observation applications. However, due to the judge and weigh between image spatial resolution and image spectral resolution, the spectral resolution of PAN images is often limited. MS images have a wider bandwidth than PAN images and generally have a higher spectral resolution. Combining high quality spatial resolution panchromatic images with hyperspectral resolution multispectral images is an important technique for integrating spatial and spectral information, both of these information are contained in panchromatic and multi-spectral images (Yokoya et al., 2017; Zhu et al., 2018), which is a process that is called RS image fusion.

In recent years, three types of RS image fusion methods have been widely used, they are component substitution-based methods, multi-resolution analysis-based methods and Optimization-based methods, respectively.

The first type is CS-based methods, which convert the MS image to the appropriate transform domain. Then, a specific component representing the MS image space information is replaced with a PAN image, and the fused RS image is reconstructed through an appropriate inverse transformation. Intensity-hue-saturation (IHS)-based image fusion method (Gillespie et al., 1987; Tu et al., 2004), the Gram-Schmidt (GS) image fusion method (Laben & Brower, 2000), the principal component analysis (PCA) image fusion method (Chavez & Kwarteng ., 1989), and

✉ Wei Shi
shiweinwnu@163.com

¹ College of Geography and Environment Sciences, Northwest Normal University, Lanzhou 730070, China

² School of Automation, Northwestern Polytechnical University, Xi'an 710129, China

³ School of Computer Science, China University of Geosciences, Wuhan 430074, China

adaptive component substitution (ACS)-based image fusion method (Choi et al., 2011) are the most typical example based on the component substitution methods. Still need to pay attention to that analyzing the correlation between the PAN image and the replaced MS component has a large effect on the fusion result.

The second type based on the multi-resolution analysis methods, which generally protect spectral information during fusion rather than most classical component substitution-based methods. On the whole, multi-resolution analysis-based methods first extract the spatial structures from the PAN image by transformation method, such as Laplacian pyramid, wavelet transform, the extracted typical spatial structure information is then injected into the up-sampled MS image in order to yield a fused RS image. The RS image fusion methods wavelet transform-based (Aiazzi et al., 2002a) or curvelet transform-based (Nencini et al., 2007), the modulation transfer function (MTF)-based (Palsson et al., 2016), and the smoothing filter intensity modulation (SFIM)-based method (Liu, 2000) are the most typical example based on the multi-resolution analysis methods. However, a potential drawback with these methods is that they usually produce spatial distortion.

The third type is Optimization-based methods, which based on the RS image observation model, the fusion image is regarded as solving ill-posed problems. In generally, the RS images fusion can be to deal with by minimizing the loss function during the fusion process with the prior constraints, for example, Bayesian posterior probability (Fasbender et al., 2008), the minimum mean square error (Garzelli et al., 2008), normalized Gaussian distribution-based (Zhang et al., 2012), total variation operators (TVO) (Shen et al., 2016), and sparse representation (SR)-based fusion methods (Dian et al., 2019). Especially, the selected image fusion strategy has a profound effect on the fusion performance of Optimization-based methods, but the choice of integration strategy is usually more difficult.

The above summarizes the traditional RS image fusion methods, and finds many ways to extract and select features. In other words, researchers need to select not less than one tool to transform the RS image to extract features from the beginning in the traditional method. Then researchers also were asked to design a specific fusion rules to determine which image characteristics obtained from MS and PAN image should be injected into the fused results. Ultimately, the fused result is inversely transformed to obtain a fused image.

For RS images, in most low-level image processing tasks, such as image superresolution, deblurring, inpainting (He et al., 2015; Krizhevsky et al., 2017; Ouyang et al., 2015), in recent years CNN-based algorithms have been reached the optimum accuracies, and their processing quality is constantly improving. However, in the field of

remote sensing image fusion, the introduction of the CNN model in recent years is still limited. Inspired by the immensely distinguished performance of CNN in the RS image field, a novel remote-sensing image fusion algorithm based on MSCNN is presented. Compared with the above-mentioned three typical RS image fusion algorithms, the MSCNN method can validly extract and fuse the features of the image well without artificially designing a particularly complex fusion rule. The main contributions of this article are as follows:

First, the presented fusion algorithm utilizes a set of CNN to perform multi-scale analysis on each band of a multispectral image in order to validly extract the typical characteristics of different band of multispectral images.

Second, to prevent losing the information of the original image, the max-pooling layer of the traditional CNN is replaced with a standard convolutional layer, and the standard convolutional layer has a step size of 2.

The remainder of this article is arranged as follows. Section 2 is the basically theoretical and application of CNN for RS image fusion. Section 3 introduces the proposed MSCNN sensing image fusion method. The experimental results and corresponding subjective and objective results analysis are presented in Sect. 4. Section 5 is the conclusion and summary.

CNN Model for Remote Sensing Image Fusion

CNN Model

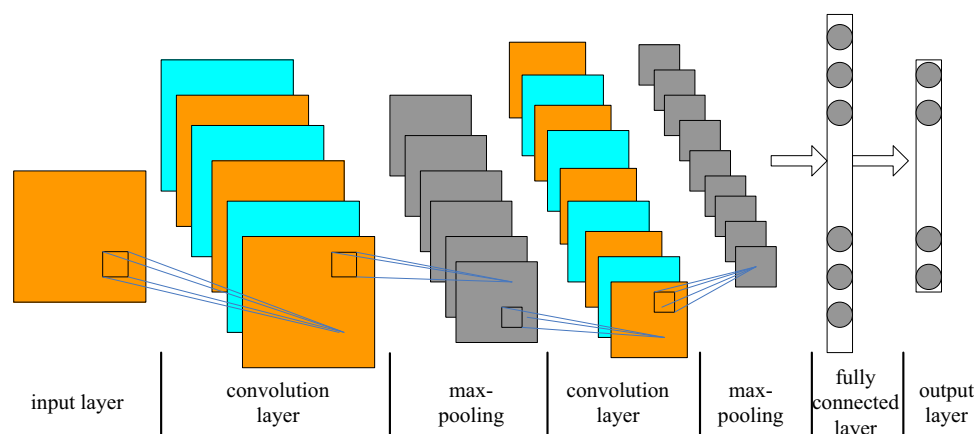
CNN is a kind of deep learning (DL) network, which mainly used to learn images at different levels of abstraction, so as to obtain multiple features of images. Figure 1 shows a classic example of CNN structure, which mainly composed of four parts.

The input information of the CNN is usually the original image X . In this paper, the characteristic map of the i -th convolutional layer of the CNN ($H_0 = X$) is represented by H_i .

$$H_i = f(H_{i-1} \otimes W_i + b_i) \quad (1)$$

where $f(\cdot)$ is the activation function, ReLU function is selected as the activation function in this paper, b_i is the bias, W_i is the convolutional kernel and the convolutional operation is represented by \otimes .

Following the convolution layer is the max-pooling layer. A fully connected network is employed to classify the extracted image features based on the input in CNN, in order to get a satisfactory probability distribution Y . CNN's basic operation is a typical mathematical model, which maps the original input image matrix to a whole new

Fig. 1 A typical CNN structure

feature expression Y through a series of dimensionality reduction or multiple levels of data transformation.

$$Y(i) = P(L = l_i | H_0; (W, b)) \quad (2)$$

Minimizing the loss function $L(W, b)$ is the ultimate goal of CNN training, the training parameters W and b of the each convolutional layer is updated layer by layer.

$$W_i = W_i - \eta \frac{\partial E(W, b)}{\partial W_i} \quad (3)$$

$$b_i = b_i - \eta \frac{\partial E(W, b)}{\partial b_i} \quad (4)$$

where, $E(W, b) = L(W, b) + \frac{\lambda}{2} W^T W$, λ is the weight decay parameter, η is the parameter of learning rate (https 2017).

RS Image Fusion Based on CNN

Most existing RS image fusion methods usually consist of two components: image characteristic extraction and image characteristic fusion. When the SR or multi-resolution analysis is used for image fusion purposes, step one is to represent the input image through a string of appropriate atoms or base filters in the dictionary obtained by learning. After deriving the expression, step two is to select the appropriate strategy to fuse the expression of the source image, so as to generate the expression of the fused image. It is worth noting that all processes can also be balanced to accomplish image characteristic extraction and image characteristic fusion using different convolution kernels. Therefore, because the convolution layer can reach the same effect as the traditional RS image fusion method, it is more reasonable to obtain the fused RS image through extract the features of different RS images by using CNN.

The CNN is employed to classify images is currently a popular practice (Krizhevsky et al., 2017). The output of CNN is the distribution probability of each category of features in RS image belonging to each category by putting a source image into the networks. The designed CNN does

not have a max-pooling layer when using CNN to process image super-resolution reconstruction problems. The output image of CNN is a reconstructed image whose size is the same as the size of the input image. In particular, labels and inputs for CNN training are high-resolution images and low-resolution images, respectively (Dong et al., 2016; Kim et al., 2016). To reduce the difference between the output of network and label, the network will continue to learn the parameters suitable for label. To achieve the purpose of fusion RS image using CNN, the same consideration from the field of RS image super-resolution reconstruction is employed. The purpose of RS image fusion is fuse PAN images with MS images to produce an image with both spectral resolutions and high spatial.

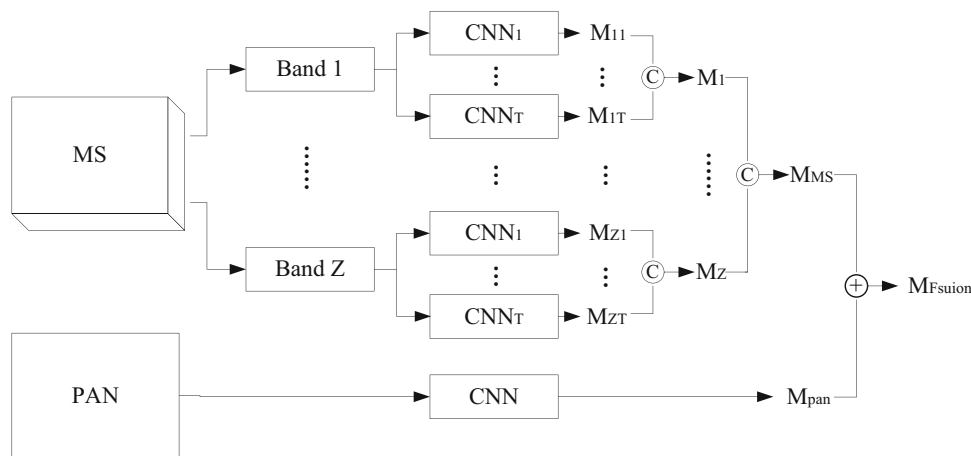
Multi-Scale CNN Method

Algorithm Execution

The conceptual workflow of the proposed RS fusion algorithm is presented in Fig. 2. The number of spectral bands is indicated by Z in the multi-spectral source image, and the number of CNN per band is represented by T .

Some similarities will be found by compare convolutional operations with traditional RS image fusion algorithms. For the method based on SR, the image is first decomposed into sparse coefficients according to the dictionary, and these sparse coefficients are considered to be characteristic of the subsequent fusion process. While in the MSCNN method shown in Fig. 2, the characteristics of RS images are extracted through a string of different convolutional kernels. These convolutional kernels are regarded as the dictionary of SR, the characteristic maps obtained by the network can regard as an expression through the dictionary mentioned above. In the image fusion process based on SR algorithms, it is often necessary to artificial design image fusion rules to determine which

Fig. 2 The proposed MSCNN fusion method



sparse coefficients in the input image require further fusion. In the proposed method, convolutional kernels are employed to fuse extracted features, and during the CNN training all the convolutional kernels are automatically updated in order to yield state-of-the-art fusion results (Shao & Cai, 2018).

Multi-Scale Analysis

Set I to be the input RS image, so $I = \{p(x,y) : 1 \leq x \leq X, 1 \leq y \leq Y\}$, where $p(x,y)$ is the pixel value of (x,y) in the source RS image I , with $X \times Y$ image spatial resolution. Suppose there is a patch $P(x,y)$, the pixel (x,y) is surrounded by the $w \times w$ window, so $P(x,y)$ can be defined as

$$P(x,y) = \{p(x - \lfloor w/2 \rfloor, y - \lfloor w/2 \rfloor, \dots, p(x + \lfloor w/2 \rfloor, y + \lfloor w/2 \rfloor)\} \tag{5}$$

where $\lfloor \cdot \rfloor$ represents the floor operation. The input source RS image is broken down into a string of overlapping image blocks with plaques of different sizes firstly, such as a Gaussian pyramid (GP), which can be described as follows:

$$w_t = \begin{cases} w_b & t = T \\ 2^{T-t} * w_b & \text{otherwise} \end{cases} \tag{6}$$

where T denotes the number of convolutional neural networks per channel ($T = 3$), $w_t (t = 1, \dots, T)$ is the base patch size of $CNN_1, CNN_2, LCNN_T$. The large image blocks are dimensioned to the same size as the smallest image block ($w_b \times w_b$), and the procedure of multi-scale blocks extracted is shown in Fig. 3, the same CNN structure is used in Fig. 3.

CNN Architecture

The proposed MSCNN algorithm has three convolutional layers in the network.

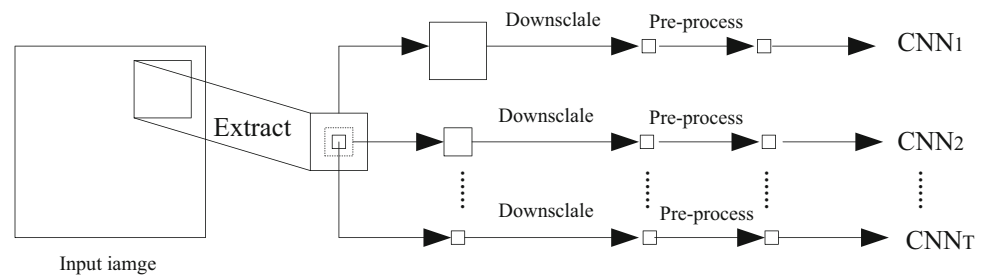
- (1) The input of the CNN is a patch of 16×16 pixels.
- (2) The first convolution layer can obtain 64 feature maps, by $3 \times 3 \times 64$ filters, and a stride of convolutional layers is set to 1.
- (3) The second convolution layer can obtain 128 feature maps, by $3 \times 3 \times 128$ filters, and a stride of convolutional layers is also set to 1.
- (4) To obtain 256 feature maps, the size of filters of another convolution layer is set to $3 \times 3 \times 256$
- (5) The 256 feature maps are forwarded to be fully connected.
- (6) The output of the CNN is a feature map.

Network Training

In this section, a training program whose purpose is to find the best parameters that fully express the entire network is presented. Let x_1 and x_2 indicate one pair of down sampled multi-spectral and panchromatic images, respectively. y is employed to represents the label of original MS image. $\{x_1^i, x_2^i, y^i\}_{i=1}^N$ is a training set, where the number of samples is represented by N . To obtain a function $f : \hat{y} = f(x_1, x_2)$, the training program was executed in the CNN, in which the predicted high-resolution multi-spectral image is represented by \hat{y} . In order to evaluate the differences between predicted results and labels correctly, the mean squared error (MSE) is employed as the loss function in the ordinary way:

$$L = \frac{1}{n} \sum_{i=1}^n \|y^i - f(x_1^i, x_2^i)\|^2 \tag{7}$$

Fig. 3 The procedure of multi-scale blocks extracted



where y^i is the high-resolution multi-spectral image, n is the batch size and $f(x_1^i, x_2^i)$ is the predicted result.

Similar to the task based on CNN (Farfade et al., 2015; Long et al., 2015), the soft max loss function is used as the objective function of the proposed CNN network. Researchers minimize the soft max loss function by selecting stochastic gradient descent (SDG) in this paper. The momentum is set to 0.9 and the weight decay is set to 0.0005. The weights in the CNN are renewed stepwise by the following equation:

$$v_{i+1} = 0.9 * v_i - 0.0005 * \theta * w_i - \theta * \frac{\partial L}{\partial w_i}, \quad \partial w_{i+1} = \partial w_i + v_{i+1} \quad (8)$$

where v, θ, i and L is the momentum parameter, learning rate parameter, iteration index parameter, and loss function, respectively. $\frac{\partial L}{\partial w_i}$ is the derivative of the loss function at w_i . The CNN framework proposed in this paper employs the popular DL framework Caffe (Jia et al., 2014). The Xavier algorithm is used to initialize the parameters used by CNN in this paper (Glorot & Bengio, 2010). The biases are initialized to 0 in every convolutional layer. The learning rate is equal and is initialized to 0.0001 in all of the convolutional layers. The learning rate dropped once during the entire training process.

Experiment Results and Analysis

Experimental Setting

The validity of the proposed algorithm is effectively evaluated through RS images from the IKONOS and Quick-Bird satellites in this section. These compared fusion methods include three CS-based methods, i.e., Gram Schmidt (GS) (Laben et al., 2000), Principal Component Analysis (PCA) (Chavez & Kwarteng, 1989), Nonlinear Intensity Hue Saturation (NIHS) (Ghahremani & Ghassemian, 2016); two MRA based algorithms, i.e., Generalized Laplacian Pyramid (GLP) (Aiazzi et al. 2002b), Additive Wavelet Luminance Proportional (AWLP) (Otazu et al., 2005); and one regularization based methods, i.e., Coupled

Nonnegative Matrix Factorization (CNMF) (Yokoya et al., 2012); two CNN-based method, i.e., Convolutional Neural Network (CNN) (Zhong et al., 2016), Deep Convolutional Neural Network (DCNN) (Shao & Cai, 2018). The original application source codes of some compared RS image fusion algorithms in this section can be downloaded free of charge on the website.¹ Specifically, the number of bands set to 4 for IKONOS and Quick-bird datasets (i.e., $z = 4$).

Considering that images acquired by different satellites have different characteristics, in order to fully evaluate the method, they are divided into two categories according to the preparation process of the test images: (1) the original multi-spectral images are considered the referenced high-resolution multi-spectral images. The low frequency analog multi-spectral and panchromatic images get ready ahead of time through Wald's protocol (Wald et al., 1997). The original multi-spectral and panchromatic images are simultaneously down sampled through a scale factor, which is set as 4. Such test data are addressed as simulated data. (2) The origin multi-spectral and panchromatic images are directly employed as inputs, such test data are addressed as real data.

The spectral quality of RS image fusion results is a very important evaluation criterion, but it is difficult to visually evaluate the spectral quality. In order to compare the spectral feature and spatial characteristics of different fusion algorithms for remote sensing images, there are several indices, including the ERGAS (Wald, 2000), SAM (Alparone et al., 2007), PSNR, UIQI (Wang & Bovik., 2002), CC, and RMSE, are used as the objective evaluation index of RS image fusion performance of different methods. Specifically, ERGAS, SAM, PSNR, UIQI, RMSE, and CC are employed to evaluate the RS image fused results based on analog data. In particular, the larger UIQI, CC and the smaller SAM, ERGAS, RMSE are shown, the better performance of fused image is obtained. In order to fair comparisons and unified computing of metrics, all computed images are normalized to the range [0, 1].

¹ For PCA, GS, AWLP, GLP and CNMF methods, the source codes can be download from: <http://openremotesensing.net/knowledgebase/a-critical-comparison-among-pansharpening-algorithms/>.

Fig. 4 Fused results on the first simulated “Dashan” image

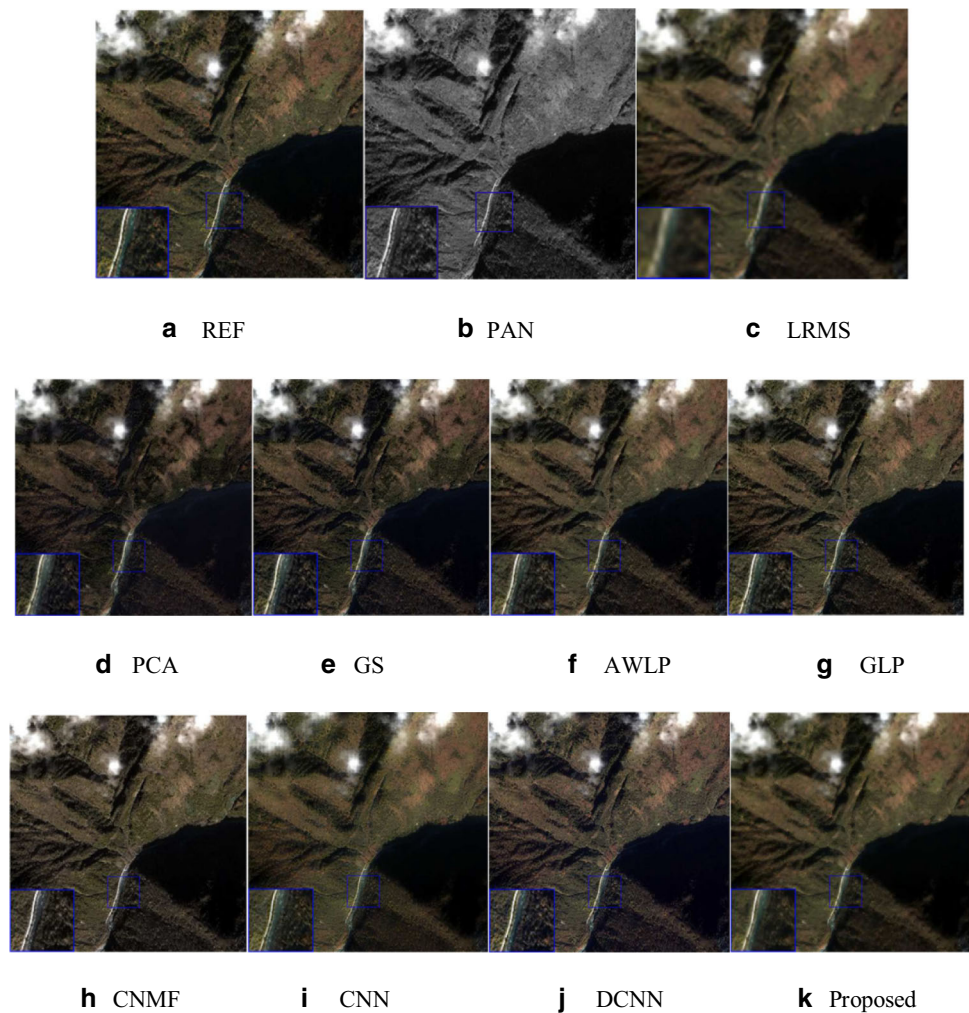


Table 1 Quantitative results on IKONOS dataset (Dashan) for different methods

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
PCA	5.9192	8.5064	22.1698	0.6736	0.1723	0.7738
GS	5.0989	5.2288	25.2566	0.6904	0.1654	0.7923
AWLP	6.1350	5.1501	25.6254	0.7218	0.1587	0.8319
GLP	6.8087	5.3411	25.3338	0.7196	0.1638	0.8296
CNMF	4.5503	2.3224	29.8108	0.8873	0.0857	0.9436
CNN	4.4296	2.4682	29.7341	0.8763	0.0903	0.9457
DCNN	4.3864	2.3815	30.0095	0.8912	0.0835	0.9582
Proposed	4.0823	2.3238	30.1428	0.9034	0.0748	0.9624
Ideal value	0	0	$+\infty$	1	0	1

Comparisons Based on Simulate Data

Firstly, we take a simulated data named “Dashan”, which collected from the IKONOS sensor to evaluate different RS image fusion algorithms, which data can be downloaded

from the website.² Figure 4a–c is the reference MS image (REF), the simulated PAN image and the simulated low-spatial resolution MS image (LRMS), respectively. Figure 4d–k displays the fused image results of PCA, GS, AWLP, GLP, CNMF, CNN, DCNN, and the proposed MSCNN method.

By comparing the fusion images displayed in Fig. 4 in detail, although it can be seen that all fusion algorithms can improve the space resolution of MS to a greater or lesser extent, some of fusion methods lead to undesirable effects such as severe spectral distortion. Although the PCA and GS methods can improve the space resolution of MS images effectively, they cause the fused image to darken as a whole. The AWLP and GLP methods significantly increase the space resolution of the MS while also providing severe spectral distortion to the fusion results. As can be seen from Fig. 4(h), there causes ringing effect and spectral distortion in the CNMF method at the same time. As can be seen from the enlarged area in the lower left

² <http://glcf.umd.edu/data/ikonos/>.

Fig. 5 Fused results on the secondly simulated image

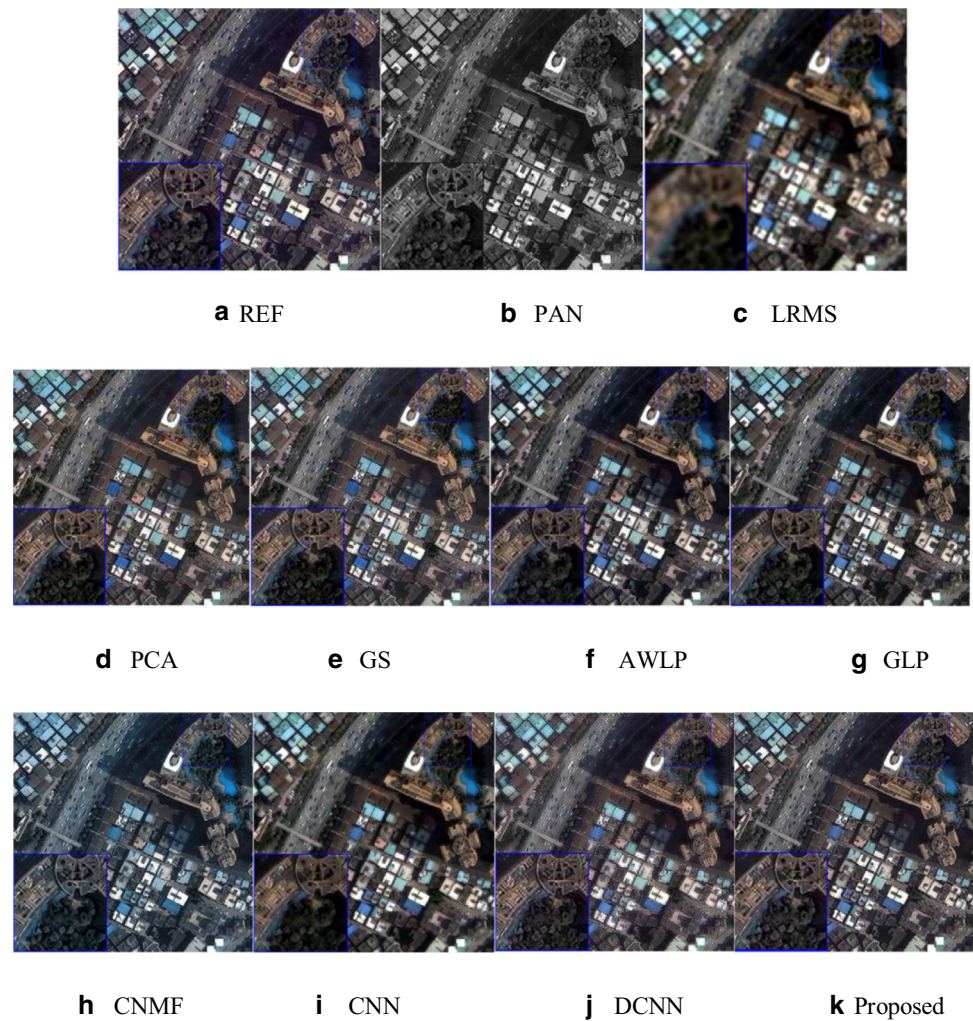


Table 2 Quantitative results on quick-bird dataset for different methods

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
PCA	2.1849	2.9649	24.9388	0.5837	0.1834	0.7527
GS	2.2152	2.9514	24.9937	0.6003	0.1769	0.7564
AWLP	1.7772	1.8851	29.6728	0.6223	0.1693	0.8174
GLP	1.6315	1.6694	30.6675	0.6186	0.1748	0.8183
CNMF	2.5169	1.6939	30.3471	0.7863	0.0963	0.9247
CNN	1.6134	1.6843	31.6935	0.8163	0.0814	0.9395
DCNN	1.5038	1.4683	32.5784	0.8618	0.0758	0.9563
Proposed	1.4827	1.2694	34.8746	0.8953	0.0649	0.9598
Ideal value	0	0	$+\infty$	1	0	1

corner of Fig. 4i and j, although the CNN and DCNN methods have no spectral distortion, the spatial resolution of MS is much less improved. The proposed method preserved spectral features from the source input multi-spectral images, which indicate the proposed method can

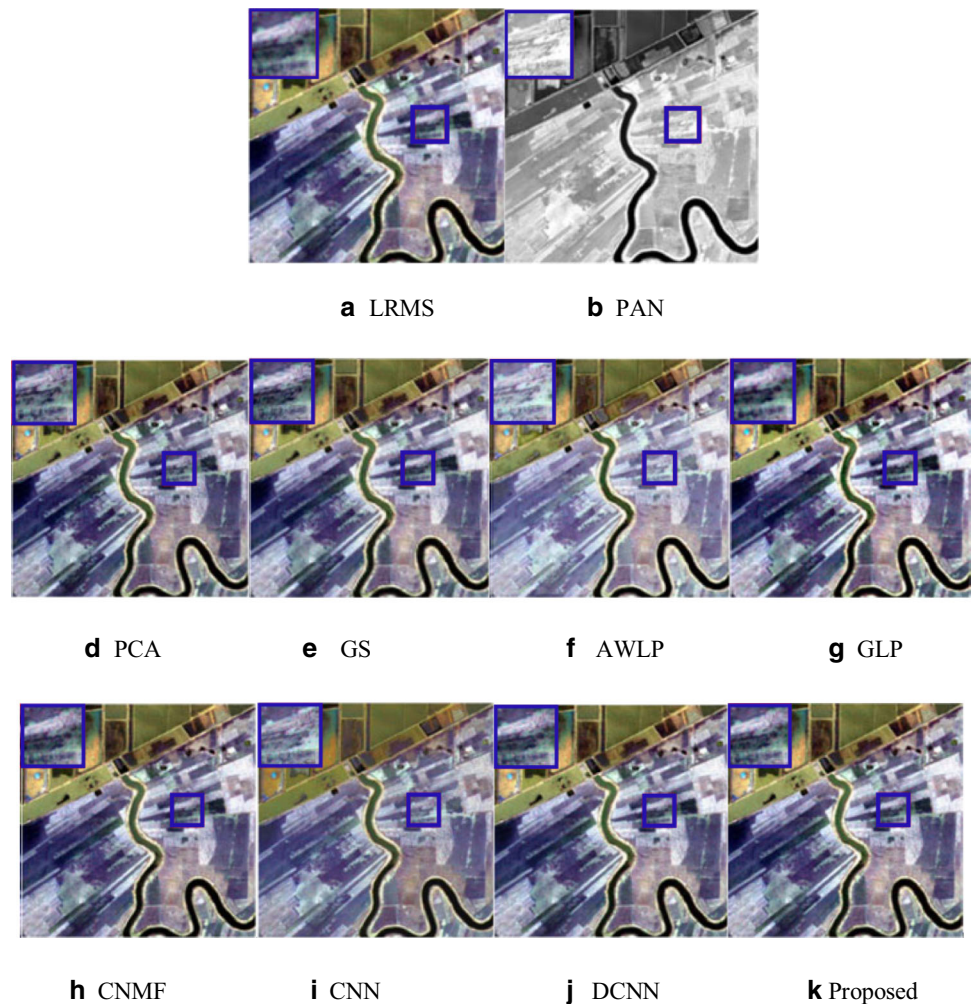
extract spectral features very effectively and can accurately inject the extracted remote sensing image features into the fusion results. Table 1 reports the objective evaluation index comparisons of fusion results (Bold is the best). From Table 1, it can be easily seen that the proposed fusion algorithm has a very competitive advantage in most indicators, except ERGAS, but the indicators of the proposed RS fusion method are very close to the indicators of the best method, which demonstrates the effectiveness of the proposed method.

Secondly, a simulated data acquired from Quick-bird sensor to evaluate different RS image fusion algorithms, which can be downloaded from the website.³ Figure 5a–c is the reference MS image (REF), the simulated PAN image and the simulated LRMS, respectively. Figure 5d–k shows the fused image results of PCA, GS, AWLP, GLP, CNMF, CNN, DCNN, and the proposed MSCNN method.

From Fig. 5, we can see that the visual results by PCA show insufficient protection of spectral information. For

³ <http://www.digitalglobe.com/product-samples>.

Fig. 6 Fused results on the real Quick-Bird data

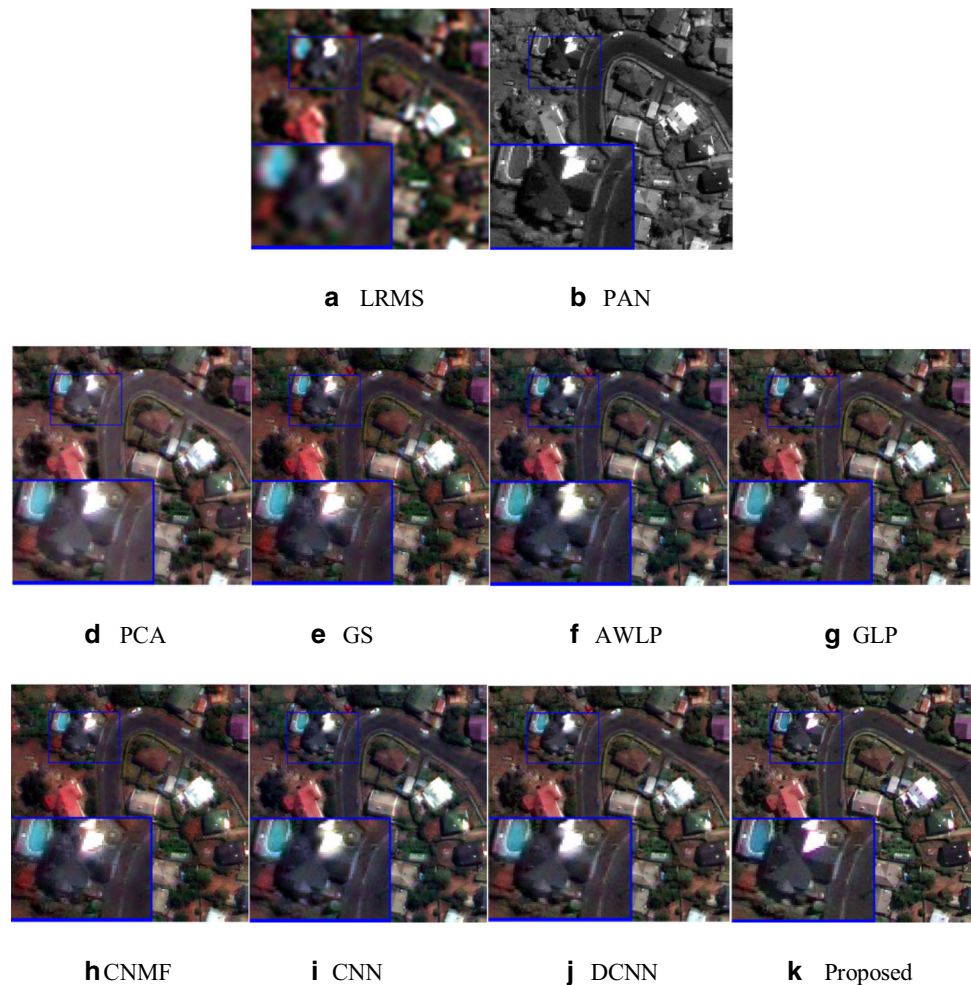


example, remote sensing images obtained by the PCA fusion method have poor spectral conservation and color contrast, which are clearly inconsistent with REF images. (i.e., Fig. 5 a). In addition, although the fused image obtained by the GS and AWLP algorithms shows satisfactory spectral preservation ability, they will lose a lot of spatial details, which may lead to poor visual effects of fused images. Although GLP approach performs excellently on preserving spectral information and increasing spatial details of the RS image, the improvement of image spatial resolution for the MS is not ideal. (e.g., see cars on the road). As shown in Fig. 5h, CNMF method causes ringing effect and spectral distortion in the fused image. A zoomed view of a particular area is provided in each fused image and is marked with a larger blue rectangle. The fusion results of CNN and DCNN are very similar to the original MS image, the spatial resolution of the DCNN is improved, and the spectral distortion is almost imperceptible, but CNN method also gives rise to the phenomenon of spectral distortion. As can be seen from the enlarged area in the lower left corner of Fig (j), although the DCNN

methods have no spectral distortion, the spatial resolution of MS is much less improved. Based on the above observations, it can be proved that the proposed method can effectively extract more spatial image information from the panchromatic to preserve spectral information, increase the visual quality of the multi-spectral image, and avoid spectral distortion. Therefore, the results of the above fusion experiments show that the method is competitive in terms of visual performance.

In order to objectively evaluate the performance of the various fusion methods involved in this paper, corresponding to the results of the RS fusion shown in Fig. 5, the evaluation index values are shown in Table 2, respectively. From the quantitative evaluation in Table 2 (in each table the best fused results are indicated in bold), we can find that in all the algorithms used in this paper for quantitative comparison, the proposed method is the most competitive method. This means that the fusion method proposed in this paper can better preserve the spectral information and inject more spatial information into the fused image in a better way.

Fig. 7 Fused results on the real-WorldView-2 data



Comparisons Based on Real Data

Each fusion method was evaluated on real data without real reference in this section. Figures 6 and 7 show the fused images on the real Quick-Bird and WorldView-2 data, respectively. The Quick-Bird data including the natural landscape and the river, and the main content of WorldView-2 data is a town. Figs 6 and 7a are interpolation images, which based on real low spatial resolution multispectral images. Figure 6 and 7b are the panchromatic images accordingly. Figure 6d–k and Fig. 7d–k display the fusion results of PCA, GS, AWLP, GLP, CNMF, CNN, DCNN, and the proposed MSCNN method.

Due to the lack of a real reference image, we can only display visual results for real data sets. For better observation, a zoomed view of a particular area is provided in each fused image and is marked with a larger blue rectangle. Compare with the actual low spatial resolution multispectral image displayed in Fig. 6 and 7a, the PCA, GS, and CNMF algorithms still suffer from some severe spectral distortion in the enlarged blue rectangular area near the river. The fused RS image obtained by the GLP

method produces a ringing effect along the river bank in Quick-Bird. The fused image yielded by the AWLP method is blurry in the whole region. Although the hyperspectral quality of the fusion results obtained by the CNN and DCNN methods is satisfactory, compared to low-resolution multispectral images, it does not improve much spatial resolution. In contrast, this method achieves a good fusion effect in the case of untraceable spectral distortion and significantly improved spatial resolution. However, the detail information of fused image in MSCNN is much clear than other fusion methods. From Fig. 6 and 7, it is very powerful proof that the proposed MSCNN RS image fusion method can get very competitive visual effects. Through the experiments of these two real data, the effectiveness of the proposed method is proved.

Conclusions

CNN has proven to be an effective method in the field of image processing. However, in the field of remote sensing image fusion, the introduction of the CNN model in recent

years is still limited. Inspired by the immensely distinguished performance of CNN in the RS image field, a novel remote-sensing image fusion algorithm based on MSCNN is presented. Compared with the above-mentioned three typical RS image fusion algorithms, the MSCNN method can validly extract and fuse the features of the image well without artificially designing a particularly complex fusion rule. The most important innovation is that the proposed remote sensing image fusion method utilizes a set of convolutional neural networks (CNN) to perform multi-scale image analysis on each band of a multispectral image in order to extract the typical characteristics of different band of multispectral images. In addition, to prevent losing the information of the original image, the max-pooling layer of the traditional CNN is replaced with a standard convolutional layer, and the standard convolutional layer has one step size of 2. The RS image fusion results presented in this paper demonstrate that the proposed method is not only competitive with the most advanced methods, but also superior to other classical methods.

Acknowledgement The work of this paper was supported by the National Natural Science Foundation of China (Project Number: 41904028).

References

- Aiazzi, B., Alparone, L., Baronti, S., & Garzelli, A. (2002). Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10), 2300–2312.
- Aiazzi, B., Alparone, L., Baronti, S., & Garzelli, A. (2002). Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40, 2300–2312.
- Alparone, L., Wald, L., Chanussot, J., Thomas, C., Gamba, P., & Bruce, L. M. (2007). Comparison of pansharpening algorithms: Outcome of the 2006 GRS-S data fusion contest. *IEEE Transactions on Geoscience and Remote Sensing*, 45(10), 3012–3021.
- Chavez, P. S., Jr., & Kwarteng, A. Y. (1989). Extracting spectral contrast in landsat thematic map- per image data using selective principal component analysis. *Photogram Engineering and Remote Sensing*, 55(3), 339–348.
- Chavez, P. S., & Kwarteng, A. Y. (1989). Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis. *Photogram Engineering and Remote Sensing*, 55, 339–348.
- Choi, J., Yu, K., & Kim, Y. (2011). A new adaptive component substitution based satellite image fusion by using partial replacement. *IEEE Transactions on Geoscience and Remote Sensing*, 49(1), 295–309.
- Deng, L.-J., Feng, M., & Tai, X.-C. (2019). The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior. *Information Fusion*, 52, 76–89.
- Dian, R., Li, S., Fang, L., & Wei, Qi. (2019). Multispectral and hyperspectral image fusion with spatial-spectral sparse representation. *Information Fusion*, 49, 262–270.
- Dong, C., Loy, C. C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2), 295–307.
- Farfadi, S., Saberian, M. Li, L. (2015) Multi-view face detection using deep convolutional neural networks, in: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 643–650.
- Fasbender, D., Radoux, J., & Bogaert, P. (2008). Bayesian data fusion for adaptable image pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6), 1847–1857.
- Garzelli, A., Nencini, F., & Capobianco, L. (2008). Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1), 228–236.
- Ghahremani, M., & Ghassemian, H. (2016). Nonlinear IHS: a Promising method for pansharpening. *IEEE Geoscience and Remote Sensing Letters*, 13, 1606–1610.
- Gillespie, A. R., Kahle, A. B., & Walker, R. E. (1987). Color enhancement of highly correlated images. 2. channel ratio and chromaticity transformation techniques. *Remote Sensing Environment*, 22, 345–365.
- Glorot, X. Bengio, Y. (2010) Understanding the difficulty of training deep feed forward neural networks, in: International Conference on Artificial Intelligence and Statistics
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916.
- https://en.wikipedia.org/wiki/Deep_learning . 01-Jan-2017
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., Caffe: convolutional architecture for fast feature embedding, In: Proceedings of the ACM International Conference on Multimedia, 2014, pp. 675–678.
- Kim, J., Lee, J. K., and LEE, K.M. (2016) Accurate image super-resolution using very deep convolutional networks. In Las Vegas, NV, USA, 2016, pp. 1646–1654.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90.
- Laben, C.A., Brower, B.V. (2000) Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening.. US Patent 6011875.
- Laben, C. A., Brower, B. V.(2000) Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. Google Patents US 6011875 A.
- Liu, J. G. (2000). “Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details. *International Journal of Remote Sensing*, 21, 3461–3472.
- Long, J., Shelhamer, E., Darrell T. (2015) Fully convolutional networks for semantic segmentation, In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440
- Matikainen, L., Karila, K., Hyypä, J., Litkey, P., Puttonen, E., & Ahokas, E. (2017). Object-based analysis of multispectral airborne laser scanner data for landcover classification and map updating. *ISPRS Journal of Photogrammetry and Remote Sensing*, 128, 298–313.
- Nencini, F., Garzelli, A., Baronti, S., & Alparone, L. (2007). Remote sensing image fusion using the curvelet transform. *Information Fusion*, 8, 143–156.
- Otazu, X., González-Audicana, M., Fors, O., & Núñez, J. (2005). Introduction of sensor spectral response into image fusion methods. *IEEE Transactions on Geoscience and Remote Sensing*, 43, 2376–2385.

- Ouyang, W. et al. (2015) Deepid-net: Deformable deep convolutional neural networks for object detection. In Boston, MA, USA, 2015, pp. 2403–2412.
- Palsson, F., Sveinsson, J. R., Ulfarsson, M. O., & Benediktsson, J. A. (2016). MTF-based deblurring using a wiener filter for CS and MRA pansharpening methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 9, 2255–2269.
- Plowright, A. A., Coops, N. C., Chance, C. M., Sheppard, S. R. J., & Aven, N. W. (2017). Multi-scale analysis of relationship between imperviousness and urban tree height using airborne remote sensing. *Remote Sensing of Environment*, 194, 391–400.
- Posselt, R., Mueller, R. W., Stöckli, R., & Trentmann, J. (2012). Remote sensing of solar surface radiation for climate monitoring—the CM-SAF retrieval in international comparison. *Remote Sensing of Environment*, 118(6), 186–198.
- Shao, Z., & Cai, J. (2018). Remote sensing image fusion with deep convolutional neural network. *IEEE Journal Of Selected Topics In Applied Earth Observations And Remote Sensing*, 11(5), 1656–1669.
- Shen, H., Meng, X., & Zhang, L. (2016). An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12), 7135–7148.
- Tu, T. M., Huang, P. S., Hung, C. L., & Chang, C. P. (2004). A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geoscience and Remote Sensing Letters*, 1(4), 309–312.
- Wald, L. (2000) Quality of high resolution synthesised images: Is there a simple criterion?. In Proceedings 3rd conference fusion earth data: Merging point meas. raster maps remotely sensed images, Sophia-Antipolis, France, pp. 99–103.
- Wald, L., Ranchin, T., & Mangolini, M. (1997). Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images. *Photogrammetric Engineering and Remote Sensing*, 63, 691–699.
- Wang, Z., & Bovik, A. C. (2002). A universal image quality index. *IEEE Signal Processing Letters*, 9(3), 81–84.
- Yokoya, N., Yairi, T., & Iwasaki, A. (2012). Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 50, 528–537.
- Yokoya, N., Grohnfeldt, C., & Chanussot, J. (2017). Hyperspectral and multispectral data fusion: a comparative review of the recent literature. *IEEE Geoscience and Remote Sensing Magazine*, 5, 29–56.
- Zhang, L., Shen, H., Gong, W., & Zhang, H. (2012). Adjustable model-based fusion method for multispectral and panchromatic images. *IEEE Transactions on Systems, Man, and Cybernetics Part B (Cybernetics)*, 42(6), 1693–1704.
- Zhang, P., Gong, M., Su, L., Liu, J., & Li, Z. (2016). Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116, 24–41.
- Zhong, J., Yang, B., Huang, G., et al. (2016). Remote sensing image fusion with convolutional neural network. *Sensing Imaging*, 17(1), 339–348.
- Zhu, Z., Yin, H., Chai, Y., Li, Y., & Qi, G. (2018). A novel multi-modality image fusion method based on image decomposition and sparse representation. *Information Science*, 432, 516–529.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.