This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Remote Sensing Image Fusion With Deep Convolutional Neural Network

Zhenfeng Shao, *Member, IEEE*, and Jiajun Cai

***Abstract*—Remote sensing images with different spatial and spectral resolution, such as panchromatic (PAN) images and multispectral (MS) images, can be captured by many earth-observing satellites. Normally, PAN images possess high spatial resolution but low spectral resolution, while MS images have high spectral resolution with low spatial resolution. In order to integrate spatial and spectral information contained in the PAN and MS images, image fusion techniques are commonly adopted to generate remote sensing images at both high spatial and spectral resolution. In this study, based on the deep convolutional neural network, a remote sensing image fusion method that can adequately extract spectral and spatial features from source images is proposed. The major innovation of this study is that the proposed fusion method contains a two branches network with the deeper structure which can capture salient features of the MS and PAN images separately. Besides, the residual learning is adopted in our network to thoroughly study the relationship between the high- and low-resolution MS images. The proposed method mainly consists of two procedures. First, spatial and spectral features are respectively extracted from the MS and PAN images by convolutional layers with different depth. Second, the feature fusion procedure utilizes the extracted features from the former step to yield fused images. By evaluating the performance on the QuickBird and Gaofen-1 images, our proposed method provides better results compared with other classical methods.**

***Index Terms*—Deep convolutional neural network, multispectral image, panchromatic image, remote sensing image fusion.**

I. Introduction

SATELLITE sensors can detect and record the electromagnetic wave reflected by earth surfaces, while remote sensing images are carriers to store this type of information for applications, such as environment and climate monitoring [1], [2] and land cover change detection and classification [3], [4]. Many earth-observing satellites, such as Landsat, IKONOS, Gaofen-1, QuickBird, can simultaneously shoot a panchromatic image and a multispectral image within the same coverage areas. Since the reflectance value varies by land covers and spectral bands, multispectral (MS) images can record more information on earth surfaces than panchromatic (PAN) images. However, given the signal-noise ratio and tradeoffs of sensors, the spatial resolution of MS images is usually lower than that of PAN images. Image fusion is thus designed to make the most of spatial and spectral information by fusing coregistered PAN images and MS images. Ideally, the fused images should have the spatial resolution of PAN images and preserve the spectral information of MS images.

Many remote sensing image fusion methods have been proposed in recent years and can be classified into three types: component substitution, multiresolution analysis, and sparse representation.

The basic idea of component substitution is to transform MS images into another space and replace the principal component by PAN images before inversely transforming the whole dataset to the original domain. This kind of methods is extensively used because of their high computation efficiency. Commonly used implementations of this type of methods include intensity-hue-saturation (IHS) transform [5] and principal component analysis [6]. To extend the original IHS transform to images with more than three bands, the generalized IHS (GIHS) [7] was proposed, while adaptive IHS (AIHS) [8] was presented to learn coefficients for better fusion results adaptively. Dou *et al.* [48] summarized this kind of methods and proposed a general framework to implement them. The component substitution based method is generally performed on the whole image, and thus, can be regarded as a global approach. The global method can preserve the spatial details form source images, though it probably renders severe spectral distortions.

Multiresolution analysis has been proved as a powerful tool used in various fields, including image fusion. This type of methods first decomposes each source image into a low-frequency subband and a sequence of high-frequency subbands at different scales and directions. Then, fusion rules are selected according to characteristics of corresponding subbands so that redundancy and complementary information of each image can be merged. The inverse transform is performed on fused subbands to yield the fusion data at the final stage. Common multiresolution analysis tools include Laplacian pyramid transform [9], wavelet transform [10], curvelet transform, second generation curvelet transform [11], [12], and nonsubsampled contourlet transform

Manuscript received November 27, 2017; revised January 11, 2018; accepted February 6, 2018. This work was supported by the National Key R&D Plan on Strategic International Scientific and Technological Innovation Cooperation Special Project 2016YFE0202300, the Fundamental Research Funds for the Central Universities (2042016kf0179 and 2042016kf1019), the Wuhan Chen Guang Project (2016070204010114), the Guangzhou Science and Technology Project (201604020070), the Special Task of Technical Innovation in Hubei Province (2016AAA018), and the National Natural Science Foundation of China under Grant 61671332, Grant 41771452, and Grant 41771454. *(Corresponding author: Jiajun Cai.)*

The authors are with the State Key Laboratory for Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China, and also with the Collaborative Innovation Center for Geospatial Technology, Wuhan 430079, China (e-mail: shaozhenfeng@whu.edu.cn; cai_jiajun@foxmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTARS.2018.2805923

1939-1404 © 2018 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[13]. The multiresolution analysis methods can decompose source images into a series of subbands which contain diverse detailed information. The decomposed images also simulate the human vision model to make it easier to interpret the fusion results [16]. Besides, the adopted fusion rules also influence the final fusion results more or less. In fact, most fusion methods based on multiresolution analysis strengthen the fusion rules according to different fusion targets so that better fusion results are obtained [17], [18]. Methods based on multiresolution analysis can accurately extract features from the decomposed images at different scales, thus reducing halo and aliasing artifacts in the fusion process.

Sparse representation has been a hot research topic in recent years and has shown good performance in remote sensing image fusion. This kind of methods assumes all of the image patches as a linear combination of specific dictionary atoms [19]–[21]. Yang [22] and Li [23] first proposed the image fusion methods based on sparse representation, and they considered the sparse coefficient as a significant measure of images. Ding *et al.* [24] proposed infrared and visible images fusion based on sparse representation, which fully discussed the influence of different overcomplete dictionaries used for sparse representation on fusion performance. For multispectral and panchromatic images, Li [25] proposed their methods to construct a dictionary of unknown high-resolution MS images and used this dictionary to yield fusion results. Wei [26] also adopted sparse representation technique to design a novel sparse regularization term which aimed to achieve hyperspectral and multispectral images fusion. The main steps of image fusion methods based on sparse representation can be described as follows: first, source images are divided into image patches according to a specific sliding distance, and then these patches are transformed into vectors via lexicographic ordering. Afterwards, sparse coefficients can be obtained via linear expressions of vectorized patches on a predefined dictionary. By fusing sparse coefficients from different source images according to a specific rule, such as the max-activity level rule [27], the fused sparse representation can be obtained. Finally, fused image patches can be derived by multiplying dictionary and fused coefficients and are integrated together as the final fusion results. Particularly, the fusion performance of sparse representation based methods is influenced by the selected fusion strategies.

Summarizing the traditional remote sensing image fusion methods, we can find there are many procedures involving extracting and choosing features. In other words, in traditional methods, we need to select one or more tools to transform images in order to extract features at the start. Then, we also need to design specific rules to decide which features from source images should be injected to fusion result. Finally, the fusion result needs to be inverse-transformed to obtain fusion image. Inspired by the outstanding performance of deep learning in many different fields and limited usage of this technique in the remote sensing image fusion field, we hope to find a better solution for these procedures by adopting deep learning perspective. Therefore, a new remote sensing image fusion framework with two branches is proposed in this paper. This framework is mainly based on the convolutional neural network (CNN), of which the depth is extended for achieving better fusion results. The fusion network consists of two branches and a main thread. Two branches are utilized to extract features from the MS and PAN images, while the main thread fuses the extracted features from branches to yield final results. Compared to the three types of fusion methods mentioned above, the proposed framework can simultaneously extract and fuse features and does not involve any manually designed fusion rules.

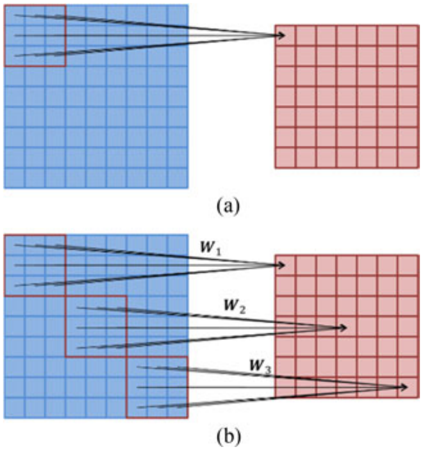


Fig. 1. Sketch map of local receptive fields and shared weights of the convolutional layer. (a) Local receptive fields. (b) Shared weights.

The rest of this paper is organized as follows. Section II is the background of CNN model and its usage for remote sensing image fusion. The proposed fusion network design is presented in Section III. Section IV includes experimental results and corresponding analysis. Concluding remarks are summarized in Section V.

II. CNN Model for Remote Sensing Image Fusion

A. *CNN Model*

CNN [28], [30], [41] is one of the most popular networks developed in recent years because of its excellent performance in image classification [29], [30], [49], target detection [31], [32], face recognition [33], and pedestrian detection [34]. What is more, the CNN has been successfully applied to remote sensing field to solve hyperspectral image classification problems [50]–[53]. CNN achieves properties of the shift, scale, and distortion invariant by fusing three basic architectural ideas, which include local receptive fields, shared weights, and subsampling. First, local receptive fields indicate each neuron at a certain convolutional layer will be connected to only a spatially neighboring region of its previous layer, which helps the network extract primary visual characteristics. In other words, neurons in convolutional layer l take a subset of neurons in layer $l - 1$ as input. Second, shared weights mean the weight of a convolutional kernel keeps the same when it is used to generate a feature map at a certain layer. As a result, the number of parameters required to be trained in the CNN will dramatically decrease compared with that in conventional neural networks. Fig. 1 shows the local receptive fields and shared weights, where $\boldsymbol{W}_1 = \boldsymbol{W}_2 = \boldsymbol{W}_3$ in Fig. 1(b). Third, subsampling lowers the spatial resolution of a feature map, which combines with convolution operator to achieve translation invariance.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Let $x^{(i)}$ and $y^{(j)}$ denote the ith input feature map and jth output feature map of a convolutional layer. Therefore, a convolutional operation with activation function applied to $x^{(i)}$ can be expressed as

$$y^{(j)} = f\left(b^{(j)} + \sum_i k^{(i)(j)} * x^{(i)}\right) \tag{1}$$

where $k^{(i)(j)}$ is a convolutional kernel which is applied to ith input feature map to obtain the jth output feature map, and $b^{(j)}$ indicates the bias. The symbol $*$ indicates the convolutional operator and f indicates the activation function. If a convolutional layer consists of M input feature maps and N output feature maps, there will be N convolutional kernels with the size of $d \times d \times M$, where $d \times d$ also indicates the size of local receptive fields. Besides, each kernel contains a bias. Selection of proper activation function is an essential part of a neural network, and conventional activation functions include Sigmoid, Tanh, and ReLU [35]. For example, (1) can be re-expressed as the following by incorporating the nonlinear ReLU activation function:

$$y^{(j)} = \max\left(0, b^{(j)} + \sum_i k^{(i)(j)} * x^{(i)}\right). \tag{2}$$

B. Remote Sensing Image Fusion Based on CNN

Most of current remote sensing image fusion methods usually contain two steps: Feature extraction and feature fusion. For example, when the multiresolution analysis or sparse representation is used to achieve image fusion, the first step is to express the source images via a series of base filters or atoms in a dictionary. After the expressions are derived, the second step is to choose appropriate strategies, such as weights differences, to fuse expressions of source images so that expressions of the fusion image can be generated. It is noteworthy that all of the procedures can also be equalized to apply different convolutional kernels to achieve feature extraction and feature fusion. We will demonstrate this operation in Section III-A in detail. Therefore, as convolutional layers can achieve the effect as same as traditional fusion methods, it is reasonable and reliable to use CNN to extract the characteristics of different remote sensing images and fuse them to obtain fusion image.

The conventional CNN is usually adopted to solve the image classification problems [30], [54]. By putting an image into networks, the output will be the probability of the image belonging to each category. While putting CNN to deal with image super-resolution reconstruction issues, the designed CNN removes the pooling procedure, and the output of the network is reconstructed images which have the same size as input images. Specifically, inputs and labels for network training are low-resolution and high-resolution images respectively [36], [37]. In order to lower the difference between network outputs and labels, the network will continuously learn parameters to fit labels. To utilize CNN to remote sensing image fusion, we adopt the same thoughts from the field of image super-resolution reconstruction. Fusion aims to generate a MS image with high spatial resolution. Therefore, the label of the network is a high spatial resolution MS image and the inputs are a PAN image and a low spatial resolution MS image.

Presently, there are few deep learning based studies in remote sensing image fusion. Masi *et al.* [39], Palsson *et al.* [40], and Zhong *et al.* [55] directly adopted SRCNN [36], which is a popular network for image super-resolution reconstruction, to implement remote sensing image fusion. However, the first two methods only use three convolutional layers and cannot adequately leverage the depth of the network to extract deep features. Also, these methods regard the PAN images as a band and overlay it on the MS images to train the network, which ignores distinctive characteristics of these two types of images. While the third method only uses SRCNN to do image super-resolution rather than image fusion. The fusion procedure is still finished by traditional Gram-Schmidt transform method.

III. Proposed Fusion Framework

In this section, we demonstrate our remote sensing image fusion method in detail. The network architecture is shown in Fig. 2, which uses the acronym RSIFNN in the following, meaning CNN-based remote sensing image fusion.

A. Network Design

The proposed method contains the same procedures as the classical remote sensing fusion methods: Feature extraction and feature fusion. Different from other deep learning-based fusion methods [39], [40], [55], the proposed method design two branches to extract features of the MS and PAN images separately. Fig. 2 shows the whole training scheme of RSIFNN, where the input MS and PAN images generated from the QuickBird satellite are of 2.8 m and 0.7 m spatial resolution.

The output of network should be the MS image with same spatial resolution as the PAN image. This image needs to be as similar as possible to the ideal MS image (label) which obtained by a sensor with same spatial resolution as the PAN image. However, this ideal MS image does not exist. It will cause troubles for our training and performance assessment if there do not exist labels. Fortunately, this problem can be solved by using Wald's protocol [38]. In Wald's protocol, the original MS images are regarded as labels. For keeping the same spatial resolution as the original MS image (labels), the original PAN image needs to be down-sampled according to the ratio between the resolution of MS image and PAN image. At the same time, the input MS image with low spatial resolution is generated by down-sampling the original MS image and interpolating the down-sampled data. Fig. 3 shows the procedure of preparing training samples. In the case of QuickBird images, we have MS images and PAN images of 2.8 m and 0.7 m spatial resolution. The inputs are obtained by down-sampling the original MS and PAN images by factor 4 (the multiple relationships between the original MS images and PAN images) to get MS images and PAN images of 11.2 m and 2.8 m spatial resolution. Therefore, we can transform original fusion task to obtain fused MS images (2.8 m) by fusing low-resolution MS images (11.2 m) and PAN images (2.8 m). In this way, we have labels (original MS images)

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

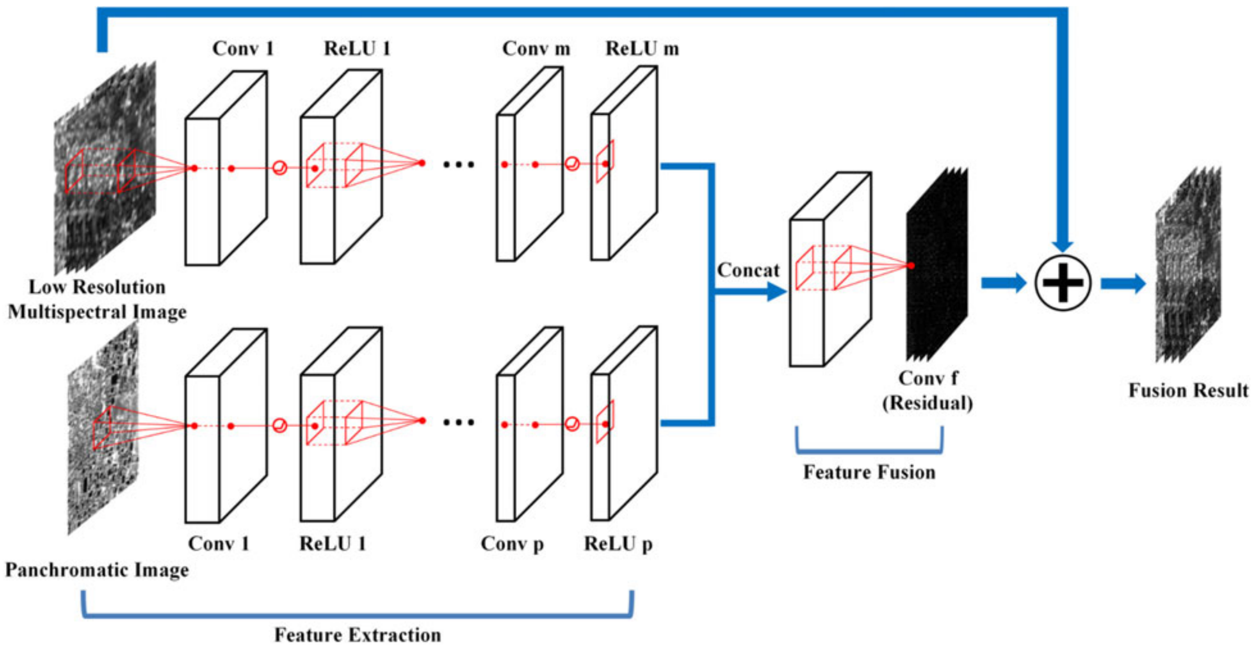


Fig. 2. Architecture of RSIFNN for remote sensing image fusion.

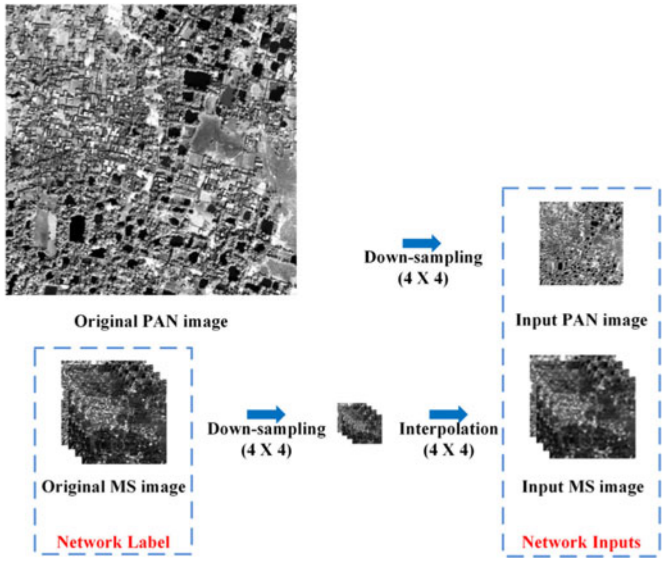


Fig. 3. Procedure of preparing network inputs.

to train our network and evaluate our fused MS images in both objective and subjective perspective.

Compared to deep learning-based fusion methods proposed in [39], [40], and [55], which only contain three convolutional layers, RSIFNN contains more layers. With the help of deeper architecture, we can exploit high nonlinearities and provide high-level features for fusion task. As shown in Fig. 2, the branch depth for feature extraction of MS images and PAN images are denoted as m and p, respectively. Except for the first layer and output layer of each branch (mth layer for the branch of MS images and pth layer for the branch of PAN images), other layers all contain 64 filters with the size of $3 \times 3 \times 64$ which means the spatial extent is 3×3 across 64 channels. Each branch also contains 64 filters at the first layer, but the size of the filter is $3 \times 3 \times k$, where k equals bands of the input image. The output layer of each branch consisting of 32 filters of the size $3 \times 3 \times 64$ exports extracted features. Then feature fusion is implemented by combining data from each branch via concatenation, and then 4 filters of the size $3 \times 3 \times 64$ are utilized to yield fusion results. The configuration of network mentioned above is summarized in Table I (suppose the MS image contains 4 bands).

TABLE I
Configuration of Each Layer

Corresponding procedure	Branch name	Layer #	Kernel size	Feature maps	Activation
Feature extraction	MS branch	1	$3 \times 3 \times 4$	64	ReLU
		m	$3 \times 3 \times 64$	32	ReLU
		others	$3 \times 3 \times 64$	64	ReLU
	PAN branch	1	$3 \times 3 \times 1$	64	ReLU
		p	$3 \times 3 \times 64$	32	ReLU
		others	$3 \times 3 \times 64$	64	ReLU
Feature fusion	Main thread	f	$3 \times 3 \times 64$	4	none

If we compare the convolutional operations to traditional fusion approaches, we can find some similarities. In case of the sparse representation-based method, images are decomposed to sparse coefficients based on the dictionary, and these sparse coefficients are regarded as features for the subsequent fusion process. While in the two branches of the network shown in Fig. 2, we extract features of images by a series of different convolutional kernels. If we regard these kernels as the dictionary, then the obtained feature maps can naturally regard as expressions via this dictionary. In the fusion process, sparse representation-based methods often involve manually designed fusion rule to decide which sparse coefficients from source images should be chosen for further fusion. In the main thread of our network, we also apply convolutional kernels to fuse extracted features, and all kernels are automatically learned during the training phase to yield better fusion results.

As suggested by the GIHS method [7], we consider that there is a mask containing necessary information between high- and low-resolution MS images. This mask is directly obtained by subtracting PAN images from MS images in GIHS. RSIFNN calculates the mask in the last layer of the network, which is named residual learning. Fusion results are obtained by overlaying the mask on the low-resolution MS image. This learning technique has been proved to work effectively in improving the learning speed and reducing memory consumption caused by the deeper structure [37].

B. Network Training

In this part, we will show the training procedure which aims to find optimal parameters to express the whole network sufficiently. Let x_1 and x_2 denote a pair of down-sampled MS and PAN images. Let y indicates the label (original MS image). Then, a training set is expressed as $\{x_1^{(i)}, x_2^{(i)}, y^{(i)}\}_{i=1}^{N}$, where N is the number of samples. The target of this training procedure is to obtain a function f: $\hat{y} = f(x_1, x_2)$, in which $\hat{y}$ is the predicted high-resolution MS image. Normally, the mean squared error is used as its loss function to measure the difference between the predicted result and label:

$$L = \frac{1}{n}\sum_{i=1}^{n}\left\| y^{(i)} - f(x_1^{(i)}, x_2^{(i)})\right\|^2 \tag{3}$$

where $y^{(i)}$ is the high-resolution MS image, $f(x_1^{(i)}, x_2^{(i)})$ is the predicted image, and n is the batch size, meaning the number of training samples which are randomly selected from the training set.

The networks proposed in [39] and [40] straightforwardly output the predicted high-resolution MS image. In other words, they directly used the network output and labels to calculate the loss function. As a matter of fact, there is a mass of redundant information between the MS images with low and high spatial resolution due to their spectral similarities. The network may not be able to extract deep features or even run out of memory because of saving too much redundancy.

In this paper, the residual learning layer can solve this problem. Because there are many similarities between low- and high-resolution MS images, we can create a residual image $r = y - x_1$ to describe their differences. Pixel values in the residual image will mostly be zero or small, and thus the whole residual image is sparse. In this way, we can ignore redundant information and just focus on the feature which mainly improves the spatial resolution for MS images. By adding the residual image to the low-resolution MS image, the high-resolution MS image is generated. The loss function now is modified as

$$L = \frac{1}{n}\sum_{i=1}^{n}\left\| r^{(i)} - g(x_1^{(i)}, x_2^{(i)})\right\|^2 \tag{4}$$

where $r^{(i)}$ is the actual residual image between low- and high-resolution MS images, and $g(x_1^{(i)}, x_2^{(i)})$ is the predicted residual image.

To realize residual learning, we can do a small modification to loss layer of the network. The loss layer consists of following parts: Input from the branch of MS image, residual image, and label (high-resolution MS image), and the predicted fusion result is generated by adding first two parts together. Then we can transform loss function (4) to (3) by the following equation:

$$\begin{aligned} L &= \frac{1}{n}\sum_{i=1}^{n}\left\| r^{(i)} - g(x_1^{(i)}, x_2^{(i)})\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\| y^{(i)} - x_1^{(i)} - g(x_1^{(i)}, x_2^{(i)})\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\| y^{(i)} - (x_1^{(i)} + g(x_1^{(i)}, x_2^{(i)}))\right\|^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\| y^{(i)} - f(x_1^{(i)}, x_2^{(i)})\right\|^2. \end{aligned} \tag{5}$$

In this way, we can still use (3) to train our network while achieving residual learning at the same time.

The mini-batch gradient descent algorithm with backpropagation [41] is used to optimize the loss function. Besides, the weights are updated by the following equation:

$$\Delta_{i+1} = \beta \cdot \Delta_i - \omega \cdot \alpha \cdot w_i^l - \alpha \cdot \frac{\partial L}{\partial w_i^l}, \quad w_{i+1}^l = w_i^l + \Delta_{i+1} \tag{6}$$

where l and i are indices of layers and iterations, α and β are the learning rate and momentum, ω is the weight decay, and $\frac{\partial L}{\partial w_i^l}$ is the derivative. These parameters are empirically set for accelerating convergence and avoiding overfitting.

It is a normal phenomenon that convolution operation will reduce the size of the input. To restrain this effect, we pad zeros before convolutions to keep the size of all feature maps maintaining same.

IV. Experiment

A. Datasets and Experimental Setting

In this section, the effectiveness of our proposed method is evaluated through images from the QuickBird[1] and Gaofen-1[2] satellites. The MS images and PAN images captured by the QuickBird satellite are of 2.8 m and 0.7 m spatial resolution, while corresponding images captured by the Gaofen-1 satellite are of 4 m and 1 m, respectively. Considering the images obtained by different satellite have different characteristics, we prepare two independent training sets for network training. The samples in each set consist of 20352 patches randomly sampled from the corresponding satellite, where 16512 patches are used for training while the rest 3840 patches are used for validating the model. The size of each patch is 33 × 33. The batch size is 128 during the training procedure. The learning rate is set to 0.0001. The momentum and weight decay are set to 0.9 and 0.0001. For weight initialization, we use the method proposed in He *et al.* [56] which is a theoretically suitable method for the network with rectified linear units. In this section, the test MS images contain four bands: Infrared (IR), Red (R),

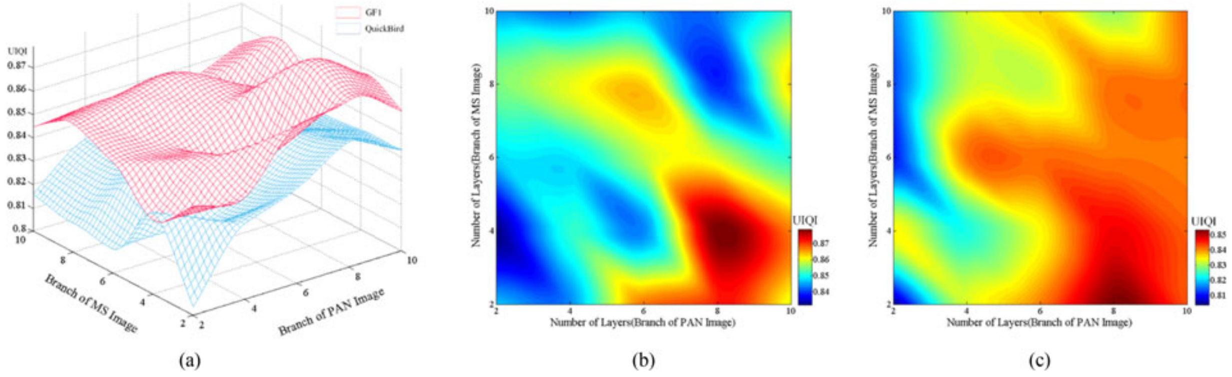


Fig. 4. UIQI values with different settings of branch depth: (a) 3-D view; (b) 2-D view of Gaofen-1 result, and (c) 2-D view of QuickBird result.

Green (G), and Blue (B). All test images are not included in training samples. Besides, all the MS and PAN images are normalized to the range [0, 1] for training and testing.

For sufficiently evaluating our proposed method, we classify our test images into two categories according to their preparing procedure. In the first category, the original MS images are regarded as the referenced high-resolution MS images (ground truth). The simulated low-resolution MS and PAN images are prepared by Wald's protocol mentioned before. Specifically, the original MS and PAN images are both down-sampled by a factor 4. We call this kind of test data as simulated data. For the second category, the origin MS and PAN images are directly used as inputs, where reference high-resolution MS images do not exist. We call this kind of test data as real data.

The adaptive IHS (AIHS) [8], wavelet transform (WT) [46], wavelet transform and sparse representation (WT+SR) [47], a nonlocal extension of BDSD (C-BDSD) [57], deep learning-based super-resolution convolutional neural network (SRCNN) [36], SRCNN combined with Gram-Schmidt (SRCNNGS) [55], and CNN-based Pan-sharpening (PNN) [39] are used as benchmark methods for comparisons.

Several indices, including the spectral angle mapper (SAM) [42], the relative global synthesis error (ERGAS) [43], the peak signal to noise ratio (PSNR), universal image quality index (UIQI) [44], root mean squared error (RMSE), correlation coefficient (CC), and quality with no reference (QNR) [45] with the spectral distortion index D_λ, and spatial distortion index D_s, are used for evaluating performances of different algorithms. Particularly, SAM, ERGAS, PSNR, UIQI, RMSE, and CC are used for evaluating the fusion results based on simulated data because references are needed for calculating these indices. As for real data, in addition to D_λ, D_s, and QNR, we also adopt SAM to evaluate fusion results. Specifically, the SAM is calculated between a down-sampled fused MS image and the low-resolution MS image. For clarification, the best values of UIQI, CC and QNR are equal to 1, while the best values of SAM, ERGAS, RMSE, D_λ, and D_s are equal to 0. The higher PSNR value indicates less noise.

B. Comparisons of Different Network Architecture

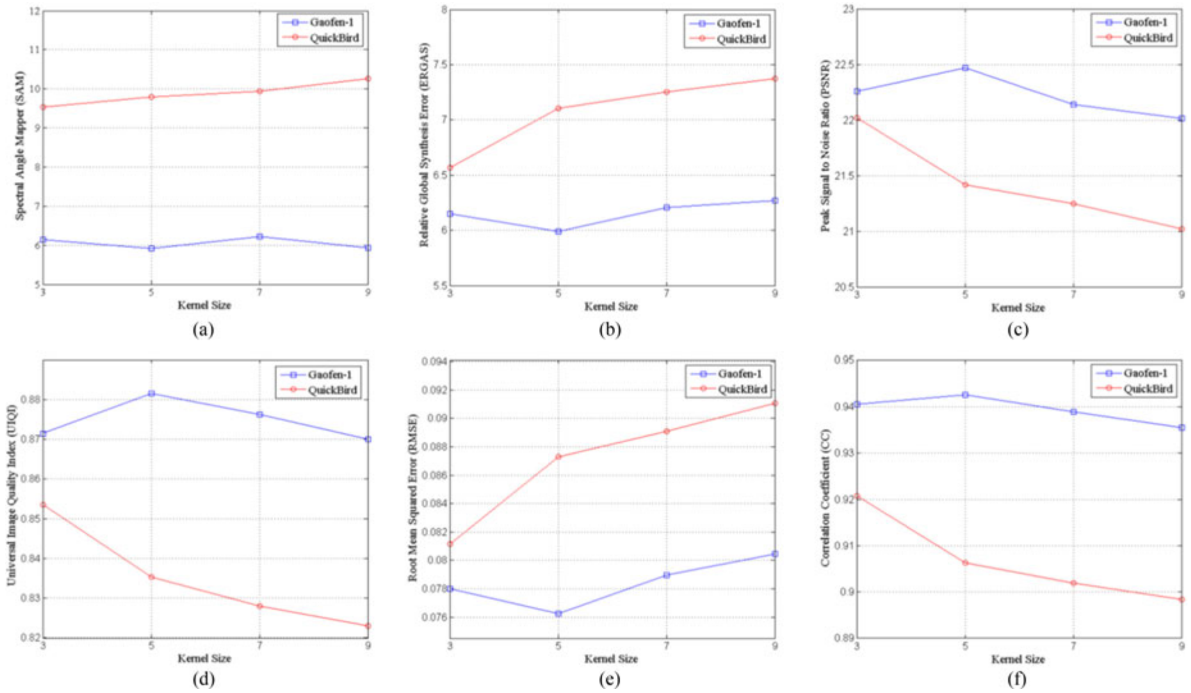
In this section, we will analyze the influence of different network architecture to fusion results. To this end, we set 50 pairs of GaoFen-1 and QuickBird simulated images for performance assessment, and all test images do not have any overlap with training samples.

The impact of the number of layers set in each branch on fusion results is analyzed in this paragraph. Here the average UIQI value based on test simulated images are used to evaluate the fusion performance. Fig. 4 shows the UIQI values with different settings of m and p. Apparently, the increase of the number of layers for feature extraction does not promise better fusion results. From the two-dimensional (2-D) view of Gaofen-1 and QuickBird results, we can realize that the best performance occurred with shallow MS branch and deep PAN branch. For the PAN image branch, eight layers are needed to extract sufficient spatial features for fusion, while two and four layers are enough for the MS image branch in QuickBird and Gaofen-1 cases respectively. In fact, low-resolution MS images have contained enough spectral information, so more spatial information from PAN images need to be injected to fusion results. Therefore, it is meaningful for the designed network to possess more layers for spatial information extraction and keep the layers for MS images shallow. As a result, the branch depth for MS and PAN images is set at two and eight in the RSIFNN, respectively.

After confirming the optimal layer number of each branch, we analyzed the impact of kernel size on fusion results. The curves of indices based on test simulated images are shown in Fig. 5. Entirely, all the indices give relatively consistent results that the optimal kernel sizes are 5 and 3 to Gaofen-1 and QuickBird respectively, which also indicates that bigger kernel size does not guarantee better fusion results. In fact, the bigger kernel size means more parameters need to be trained, which will prominently increase training time and tend to cause overfitting problem. When we scrutinize the results of Gaofen-1, it can be discovered that the results generated by 3 kernel size are better than 7 and 9, while only slightly worse than 5. Therefore, integrating the results of Gaofen-1 and QuickBird and making a tradeoff between the performances and training time, we choose 3 as the kernel size in our proposed network.

C. Comparisons Based on Simulate Data

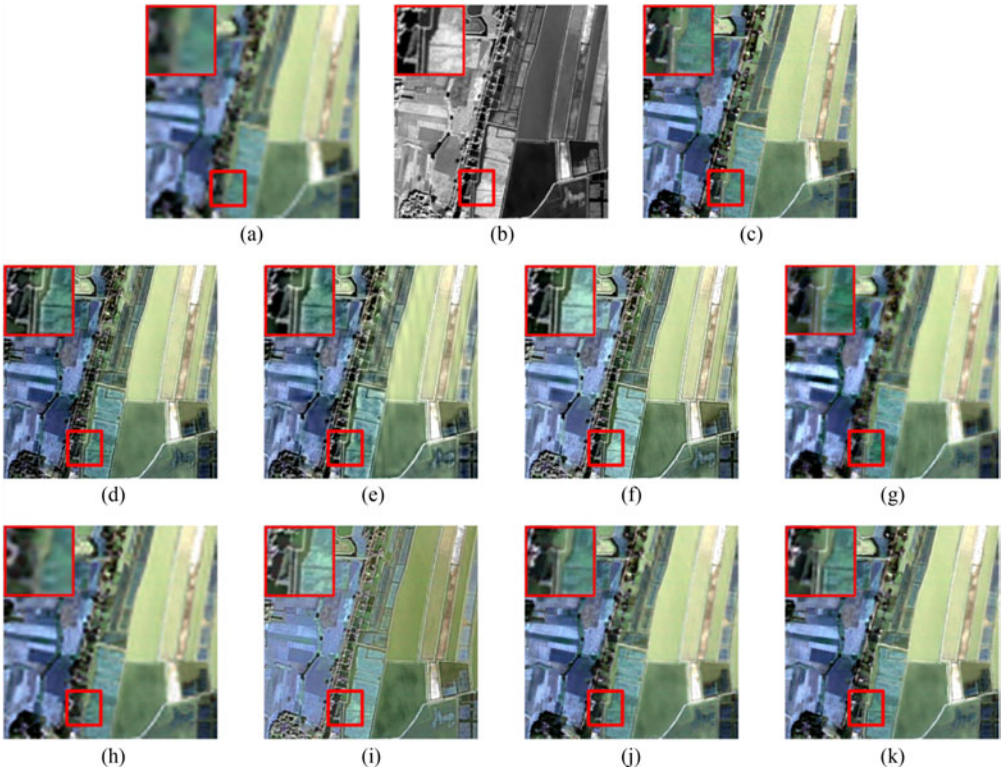
Figs. 6–9 show the fusion results on the simulated QuickBird and Gaofen-1 images. Figs. 6(a)–(c), 7(a)–(c), 8(a)–(c), and 9(a)–(c) are the simulated low spatial resolution MS image, the



(a) (b) (c)

(d) (e) (f)

Fig. 5. Values of evaluation indices with different settings of kernel size.



(a) (b) (c)

(d) (e) (f) (g)

(h) (i) (j) (k)

Fig. 6. Fused results on the first simulated QuickBird data (256×256). (a) Resampled MS image. (b) PAN image. (c) Reference MS image. (d) AIHS. (e) WT. (f) WT+SR. (g) C-BDSD. (h) SRCNN. (i) SRCNNGS. (j) PNN. (k) RSIFNN.

simulated PAN image and the reference MS image, respectively. Figs. 6(d)–(k), 7(d)–(k), 8(d)–(k), and 9(d)–(k) show the fusion results of AIHS, WT, WT+SR, C-BDSD, SRCNN, SRCNNGS, PNN, and our proposed method RSIFNN.

By inspecting these results in detail, all methods can make the MS images improve their spatial resolution more or less, but some of them contain severe spectral distortion. The AIHS method can sharply improve the spatial resolution of the MS images, but it also brings spectral distortion into fusion results. The WT method causes apparent ringing effect and spectral distortion at the same time. The WT+SR method solves the ringing effect appeared in WT method, but it still cannot restrain the

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

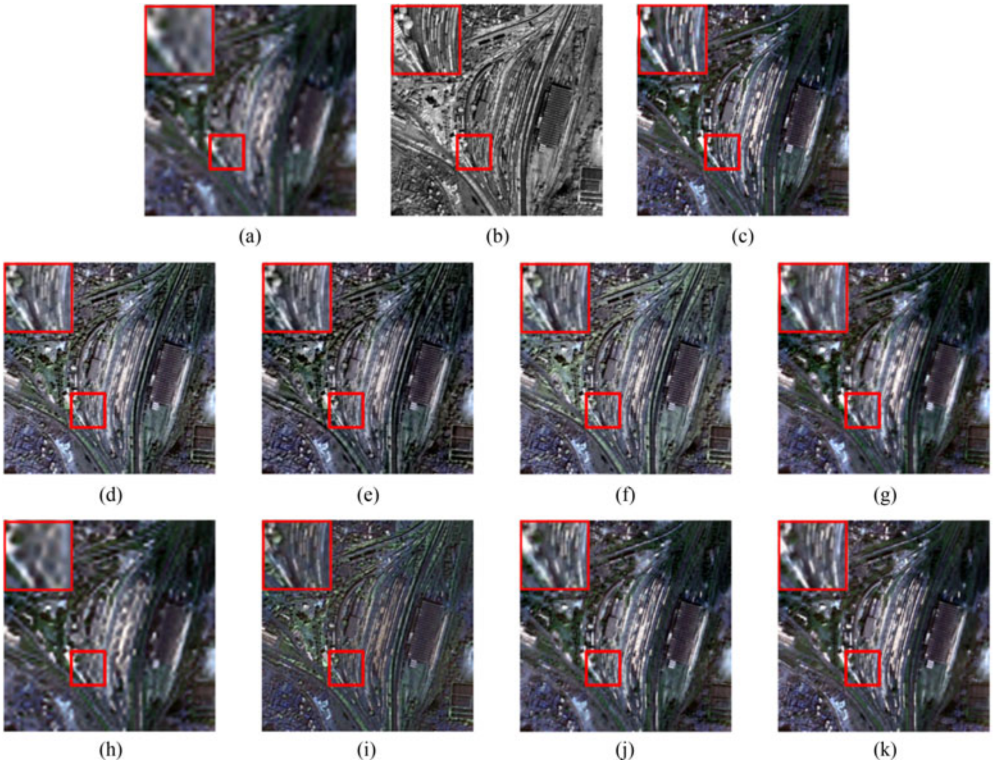


Fig. 7. Fused results on the second simulated QuickBird data (256×256). (b) PAN image. (c) Reference MS image. (d) AIHS. (e) WT. (f) WT+SR. (g) C-BDSD. (h) SRCNN. (i) SRCNNGS. (j) PNN. (k) RSIFNN.

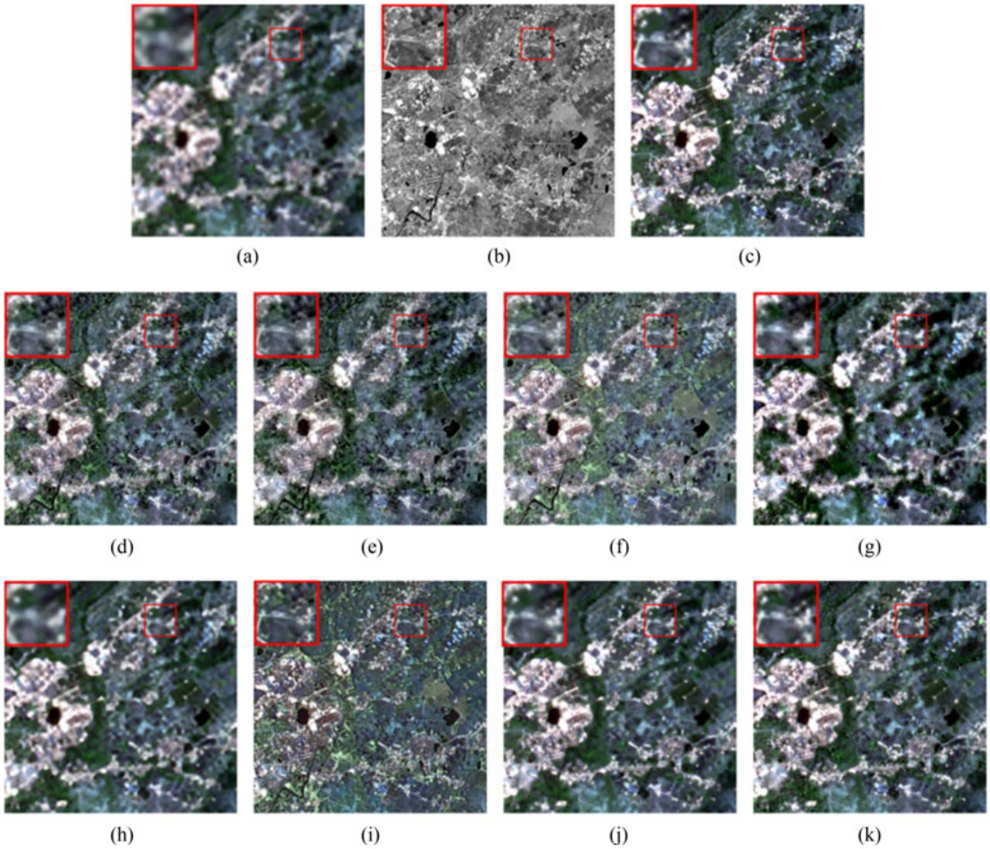


Fig. 8. Fused results on the first simulated Gaofen-1 data (300×300). (a) Resampled MS image. (b) PAN image. (c) Reference MS image. (d) AIHS. (e) WT. (f) WT+SR. (g) C-BDSD. (h) SRCNN. (i) SRCNNGS. (j) PNN. (k) RSIFNN.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

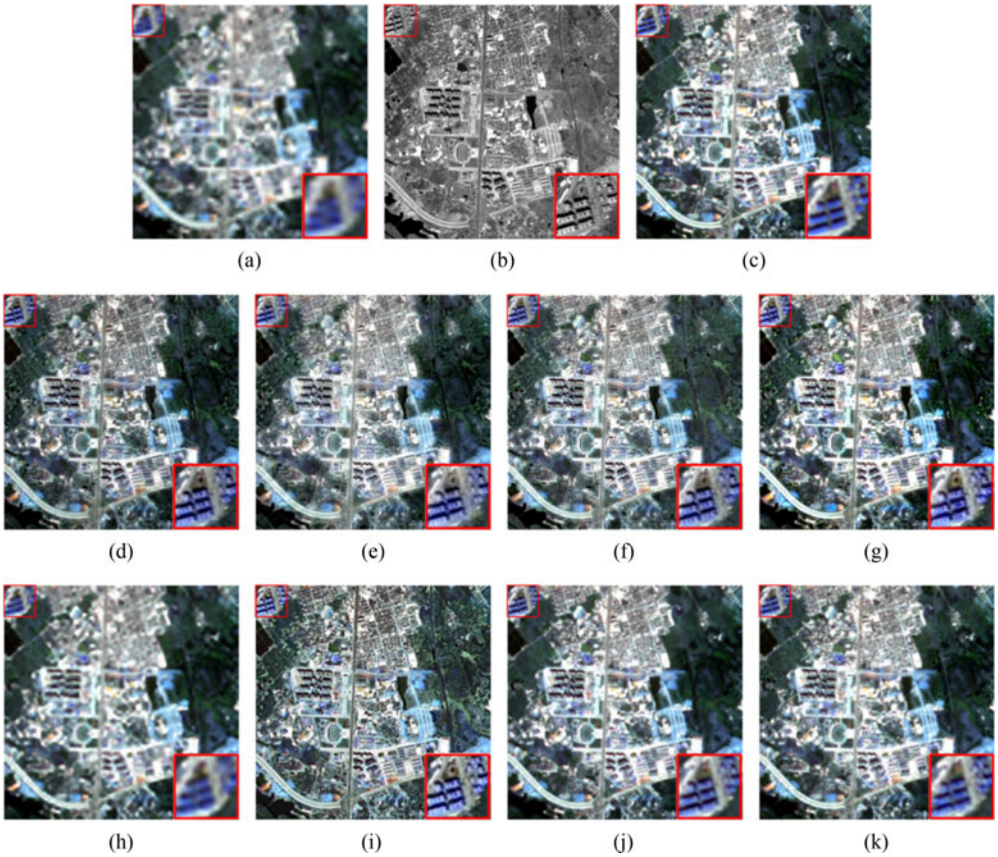


Fig. 9. Fused results on the second simulated Gaofen-1 data (300×300). (a) Resampled MS image. (b) PAN image. (c) Reference MS image. (d) AIHS, (e) WT. (f) WT+SR. (g) C-BDSD. (h) SRCNN. (i) SRCNNGS. (j) PNN. (k) RSIFNN.

color change in fusion results. The C-BDSD only suffers from a little spectral distortion, but it also causes the blurry effect in results. About four deep learning-based methods, the SRCNN method does not involve spectral distortion problem, but the improvement of spatial resolution for the MS images is minor. Although the SRCNNGS method successfully enhances the spatial resolution of the MS images, it introduces the severe spectral distortion because of adopting Gram-Schmidt method into fusion procedure. PNN and RSIFNN both preserved spectral features from the source MS images, which indicate the CNN with a shallow depth is enough for spectral feature extraction and is able to inject the extracted features to fusion results. However, the proposed method can preserve more spatial information than PNN, which can be proved by clearer linear landforms in the RSIFNN results. Besides, we can find that the results obtained by PNN or RSIFNN are much clear than SRCNN. In fact, deep learning-based image super-resolution reconstruction can only use spatial information from the input low-resolution image to generate the high-resolution result. Comparing the experiment results from PNN or PSIFNN with SRCNN, it proves that the introduction of PAN images in remote sensing image fusion can provide essential spatial information for deep learning procedure, which contributes to obtaining better fusion results.

To observe spatial differences among these methods shown in Figs. 6–9 in detail, the zoomed views of specific areas are provided in each image, marked by bigger red rectangles. The fusion results obtained by PNN and RSIFNN are very similar to the original MS images, with improved spatial resolution and barely perceptive spectral distortion. As compared to PNN, RSIFNN can generate clearer and more realistic results, which can be observed by roads in Figs. 6–8 and the color of buildings in Fig. 9. The results yielded by AIHS, WT+SR, and SRCNNGS can well fuse spatial information from the PAN images, but they also give rise to the phenomenon of spectral distortion especially in the lawn in Fig. 6. The WT method suffers from ringing effect and blurring. The C-BDSD method can produce fusion images with high spatial resolution with relatively slight spectral distortion. The SRCNN method can only generate results which are better than low spatial resolution MS images. Based on these observations, we can find RSIFNN can well extract spatial information from the PAN images to improve visual quality and preserve spectral information from the MS images to avoid spectral distortion as well. Therefore, these experiment results present the visual performance of our proposed method.

In order to evaluate the performance of each method objectively, we list the values of evaluation indices in Tables II–V, which correspond to fusion results shown in Figs. 6–9, respectively. The best results are indicated in bold in each table. Numerical results report that the better fusion results are all captured by the deep learning-based methods except SRCNNGS.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II
EVALUATION OF FUSION RESULTS USING THE FIRST SIMULATED QUICKBIRD DATA

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
AIHS	7.1920	6.8170	17.7117	0.6673	0.1303	0.8578
WT	8.5408	7.3495	17.0739	0.6110	0.1412	0.8185
WT+SR	8.1064	8.8859	15.4073	0.5570	0.1701	0.7581
C-BDSD	6.6305	5.7162	19.4371	0.7557	0.1070	0.9060
SRCNN	5.3681	4.8774	21.0041	0.7976	0.0898	0.9287
SRCNNGS	8.7743	9.3361	15.2070	0.5037	0.1789	0.6222
PNN	4.5267	3.1490	24.5109	**0.9027**	0.0605	0.9655
RSIFNN	**4.4447**	**3.1174**	**24.6136**	**0.9027**	**0.0603**	**0.9658**

TABLE III
EVALUATION OF FUSION RESULTS USING THE SECOND SIMULATED QUICKBIRD DATA

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
AIHS	12.9868	15.3532	14.7271	0.5335	0.1835	0.6718
WT	16.3843	15.2473	14.8019	0.4833	0.1820	0.6343
WT+SR	13.7732	17.6875	13.5495	0.4601	0.2103	0.5857
C-BDSD	12.8508	12.8749	16.1811	0.6313	0.1572	0.7434
SRCNN	9.8643	10.2782	18.1229	0.7082	0.1246	0.8176
SRCNNGS	12.8953	14.5804	15.3451	0.4734	0.1716	0.5956
PNN	8.9923	5.8705	23.2972	0.9051	0.0688	0.9457
RSIFNN	**8.0647**	**5.3353**	**24.1307**	**0.9235**	**0.0625**	**0.9557**

TABLE IV
EVALUATION OF FUSION RESULTS USING THE FIRST SIMULATED GAOFEN-1 DATA

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
AIHS	7.2853	8.0634	17.7622	0.7631	0.1294	0.8260
WT	9.8057	8.7855	17.0561	0.6827	0.1404	0.7927
WT+SR	8.8169	10.7731	15.4159	0.6167	0.1696	0.7486
C-BDSD	8.5604	7.8384	17.8936	0.7916	0.1284	0.8592
SRCNN	5.1311	4.7138	22.4502	0.9073	0.0756	0.9394
SRCNNGS	10.1160	11.3361	16.1134	0.5760	0.1645	0.6727
PNN	4.4557	3.3114	25.7945	0.9555	0.0516	0.9729
RSIFNN	**4.3748**	**3.1950**	**26.1229**	**0.9604**	**0.0497**	**0.9750**

TABLE V
EVALUATION OF FUSION RESULTS USING THE SECOND SIMULATED GAOFEN-1 DATA

	SAM	ERGAS	PSNR	UIQI	RMSE	CC
AIHS	5.8295	6.6835	17.7737	0.7678	0.1295	0.8810
WT	7.4052	6.0846	18.5664	0.7442	0.1183	0.8783
WT+SR	6.4707	6.5082	18.0154	0.7516	0.1259	0.8735
C-BDSD	6.3934	6.4936	18.0548	0.7676	0.1252	0.8839
SRCNN	4.3766	3.8709	22.6422	0.8982	0.0740	0.9543
SRCNNGS	8.8813	8.3545	16.9449	0.7088	0.1478	0.8326
PNN	3.6295	2.4096	26.8316	0.9630	0.0456	0.9836
RSIFNN	**3.5242**	**2.1339**	**27.8109**	**0.9688**	**0.0407**	**0.9868**

TABLE VI
EVALUATION OF FUSION RESULTS USING GAOFEN-1 AND QUICKBIRD DATASETS

		SAM	ERGAS	PSNR	UIQI	RMSE	CC
Gaofen-1	PNN	6.9228	6.6843	22.0575	0.8670	0.0797	0.9269
	RSIFNN	**6.1393**	**6.1470**	**22.2602**	**0.8715**	**0.0780**	**0.9405**
QuickBird	PNN	10.3659	7.4981	21.7947	0.8430	0.0836	0.9139
	RSIFNN	**9.5294**	**6.5648**	**22.0213**	**0.8535**	**0.0812**	**0.9207**

As we mentioned before, SRCNNGS uses Gram-Schmidt to get final fusion results, which renders the severe spectral distortion. The deep learning-based methods tend to generate results which are as close as references (labels). Therefore, under full-reference indices, the deep learning-based methods are easier to provide better performance. The results of PNN and RSIFNN are similar, but the results of RSIFNN are higher than PNN in all indices, which means that the proposed framework can inject more spatial information and preserve spectral information in a better way than PNN in the fused images.

From the quantitative evaluation in Tables II–V, we can find that PNN is the most competitive method in all methods used for comparison. Therefore, to provide a more powerful demonstration for the performance evaluation, we tested PNN and RSIFNN with another 50 pairs of Gaofen-1 and QuickBird images as mentioned in Section IV-B. The averaged results are shown in Table VI, where numerical results still show the performance of RSINFF is better than PNN.

D. Comparisons Based on Real Data

In this section, every fusion method is performed on the real data which does not have true references.

Figs. 10 and 11 show the fusion results on the real QuickBird and Gaofen-1 data respectively. The QuickBird data mainly covers a river and other natural landforms, while Gaofen-1 data mainly covers a city. Figs. 10(a) and 11(a) are interpolated images based on real low spatial resolution MS images. Figs. 10(b) and 11(b) are corresponding PAN images. The fusion results of AIHS, WT, WT+SR, C-BDSD, SRCNN, SRCNNGS, PNN, and our proposed method RSIFNN are shown in Fig. 10(c)–(j) and Fig. 11(c)–(j). For better observation, the zoomed views of specific areas are provided in each image, marked by bigger red rectangles. With the comparison of real low spatial resolution MS images shown in Figs. 10(a) and 11(a), the AIHS, WT+SR, and SRCNNGS methods still suffer from some spectral distortion in the regions of landforms near the river in Fig. 10 and roads in Fig. 11. The WT method causes ringing effect along the riverbank in QuickBird and roads in Gaofen-1. The C-BDSD method restrains the spectral distortion, but the results are blurry in the whole region. Although the SRCNN method obtains results with high spectral quality, it does not improve much spatial resolution compared to low-resolution MS images. In contrast, the PNN and RSIFNN methods both produce promising fusion results with untraceable spectral distortion and remarkable improvement of spatial resolution for MS images. However, the detail information in RSIFNN results is much clear than PNN especially shown by the texture of roundabout in zoomed view of Fig. 11.

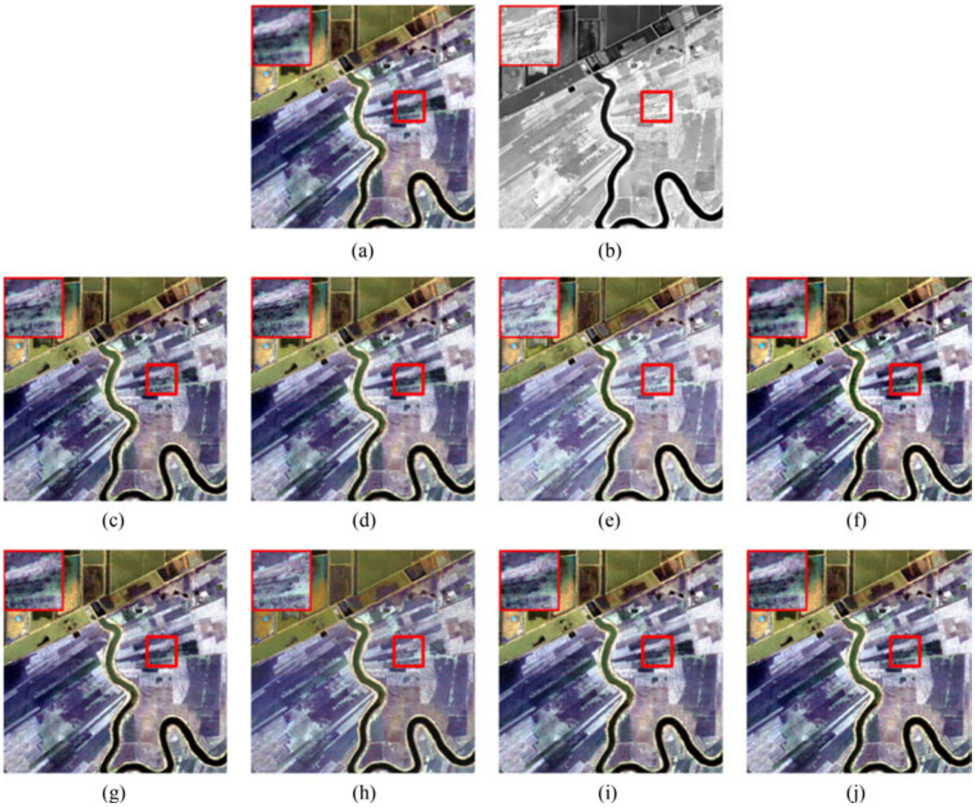


Fig. 10. Fused results on the real QuickBird data (1024×1024). (a) Resampled MS image. (b) PAN image. (c) AIHS. (d) WT. (e) WT+SR. (f) C-BDSD. (g) SRCNN. (h) SRCNNGS. (i) PNN. (j) RSIFNN.

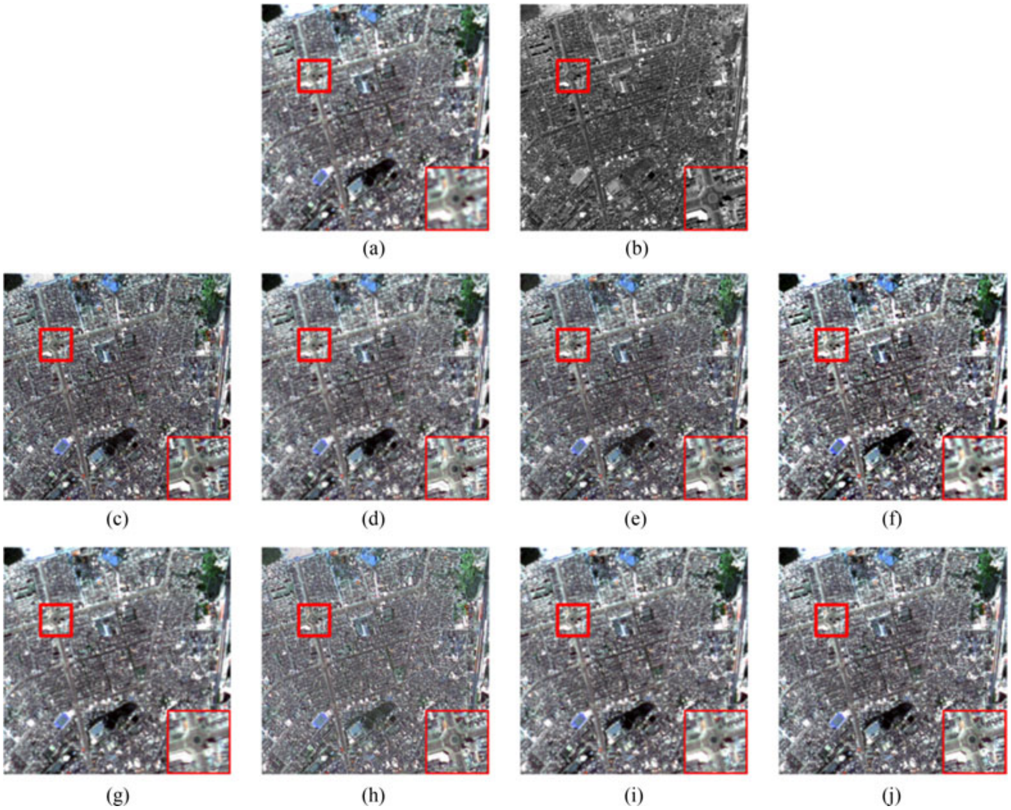


Fig. 11. Fused results on the real Gaofen-1 data (1120×1120). (a) Resampled MS image. (b) PAN image. (c) AIHS. (d) WT. (e) WT+SR. (f) C-BDSD. (g) SRCNN. (h) SRCNNGS. (i) PNN. (j) RSIFNN.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VII
EVALUATION OF FUSION RESULTS USING THE REAL QUICKBIRD DATA

	SAM	D_λ	D_s	QNR
AIHS	3.7199	0.1403	0.2677	0.6296
WT	4.6675	0.1704	0.2540	0.6189
WT+SR	7.1913	0.1338	0.3159	0.5926
C-BDSD	3.9548	**0.0105**	0.0524	0.9377
SRCNN	3.0174	0.0295	0.1796	0.7961
SRCNNGS	5.9847	0.0625	0.2540	0.6994
PNN	3.0239	0.0209	0.0540	0.9263
RSIFNN	**2.7003**	0.0277	**0.0316**	**0.9415**

TABLE VIII
EVALUATION OF FUSION RESULTS USING THE REAL GAOFEN-1 DATA

	SAM	D_λ	D_s	QNR
AIHS	3.1888	0.1087	0.3212	0.6050
WT	3.3724	0.1387	0.2292	0.6638
WT+SR	2.8309	0.1221	0.2815	0.6308
C-BDSD	2.7953	0.0128	0.0356	0.9520
SRCNN	**1.6974**	0.0105	0.1892	0.8023
SRCNNGS	3.1727	0.0550	0.3014	0.6602
PNN	2.1978	0.0582	0.0676	0.8782
RSIFNN	2.3346	**0.0053**	**0.0352**	**0.9597**

For the objective evaluation, Tables VII and VIII summarize the corresponding SAM, D_λ, D_s, and QNR values for the fusion results shown in Figs. 10 and 11. Under no-reference indices, the deep learning-based methods still generate competitive results. Overall, the proposed method produced much better results than other methods. Although the best D_λ result are obtained by C-BDSD in QuickBird images and the best SAM result are generated by SRCNN in Gaofen-1 images, the best SAM and D_λ results produced by RSIFNN with QuickBird and Gaofen-1 images indicate the proposed method can restrain spectral distortion. Therefore, these experimental results on the real QuickBird and Gaofen-1 images prove that the proposed method can keep more spectral and spatial information from source images.

E. Visualization of Residuals and Intermediate Results

As the residual learning is a significant component in RSIFNN, we also present the residuals based on real data as shown in Figs. 10 and 11. The residuals are shown in Fig. 12, and they are processed by brightness inverse for better observation. From Fig. 12, we can find that the learned residuals inject sufficient spatial information to low-resolution MS images while preserving original spectral information.

In addition to residual layer, the residual learning makes results in every layer sparse. We extracted some intermediate results based on QuickBird images utilized in Fig. 6, which is shown in Fig. 13. From these results, we can find the layers using residual learning are much sparse (most of the values are 0 or small), which contributes to reducing the memory cost.

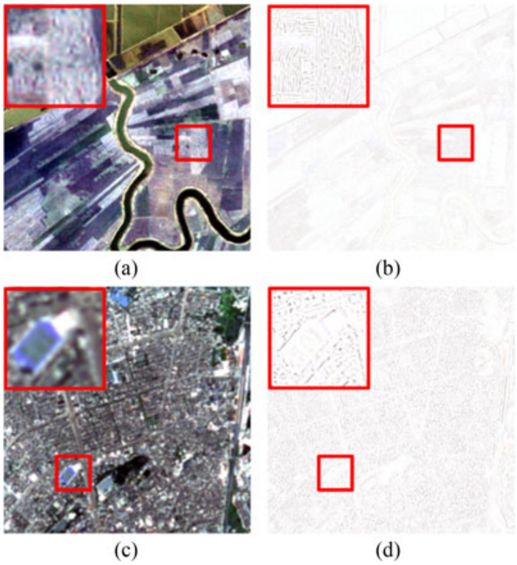


Fig. 12. Residuals of QuickBird and Gaofen-1 images. (a) and (c) Resampled MS images. (b) and (d) Residuals.

TABLE IX
RUNNING TIME (SECOND) OF THE FUSION METHODS

	Fig. 6	Fig. 7	Fig. 8	Fig. 9	Fig. 10	Fig. 11
AIHS	2.902	1.516	3.009	2.832	122.220	35.181
WT	0.196	0.186	0.197	0.195	1.440	1.531
WT+SR	14.916	15.037	20.872	20.057	310.105	344.932
C-BDSD	6.922	7.089	8.908	8.864	360.852	787.670
SRCNN	4.295	4.187	4.757	4.702	48.776	58.388
SRCNNGS	4.354	4.248	4.851	4.769	49.509	59.347
PNN	4.365	4.246	4.879	4.759	50.281	59.585
RSIFNN	0.655	0.649	1.056	0.859	10.698	11.961

F. Computational Efficiency Comparison

To clarify the computational efficiency of our proposed method, we provide some implementation details in this subsection. The training phase of the PNN and SRCNN takes about 3 h and is implemented on the GPU (NVDIA Quadro K620) through an open deep learning framework Caffe. In the same circumstance, the proposed method RSIFNN takes about 12 h for training phase, which means the training of deeper network is a much time-consuming procedure. Based on the trained parameters, the fusion procedure is performed by using an Intel.E3-1240 v5 CPU @3.50 GHz and a 16GB RAM through MATLAB R2014a. We summarize the AIHS, WT, WT+SR, C-BDSD, SRCNN, SRCNNGS, PNN, and RSIFNN methods on the images shown in Figs. 6–11, and the running time of these methods are shown in Table IX. By analyzing the data in Table IX, it is clear that the running speed of RSIFNN is only slower than WT and much quicker than PNN or SRCNN, which indicates the computational efficiency of small kernel size.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

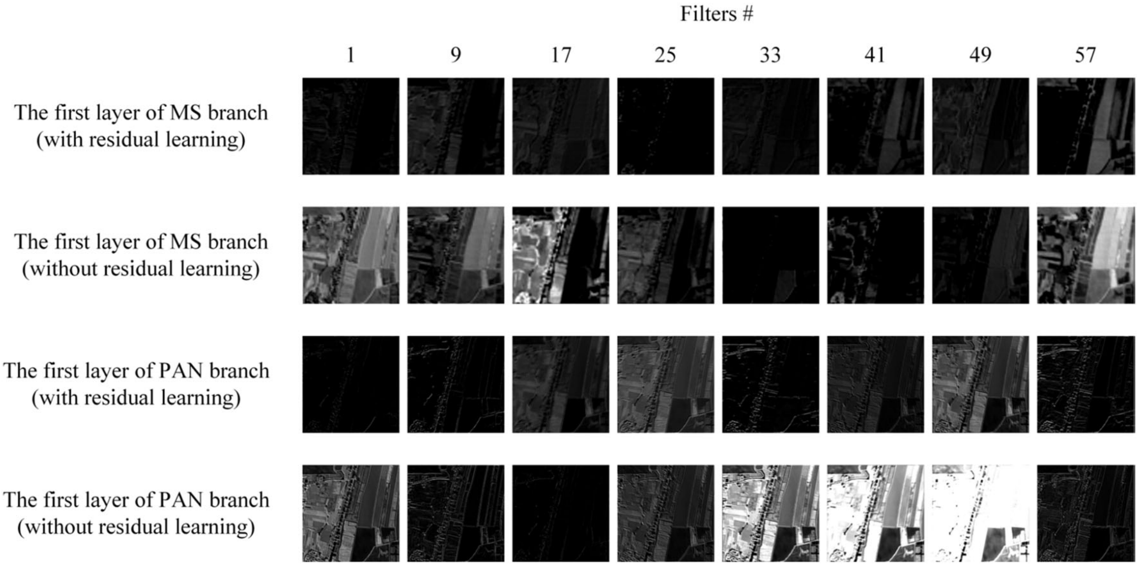

Filters #
1
9
17
25
33
41
49
57
The first layer of MS branch (with residual learning)
The first layer of MS branch (without residual learning)
The first layer of PAN branch (with residual learning)
The first layer of PAN branch (without residual learning)

Fig. 13. Intermediate results based on the QuickBird image utilized in Fig. 6.

V. Conclusion

Deep learning has been proved as a potent tool in various fields. However, few studies utilized deep learning for remote sensing image fusion. The main contribution of this paper is proposing a novel fusion framework based on the deep convolution neural network, where two branches structure and residual learning are introduced to achieve remote sensing image fusion task for the first time. Specifically, the proposed method adopts two branches with different depth to separately extract spectral and spatial features from the MS and PAN images. Also, the deeper structure in the branch of PAN images contributes to providing high-level features for the fusion process. Then, the mask between low and high spatial resolution MS images is learned to solve the fusion problem. By using residual learning layer to learn the residual image between low- and high-resolution MS images, we can ignore redundant information and just focus on the features which are essential for improving the spatial resolution for MS images. After observing experimental results, it can be concluded that our method outperforms classical fusion methods and other deep learning-based methods.

References

[1] A. A. Plowright, N. C. Coops, C. M. Chance, S. R. J. Sheppard, and N. W. Aven, "Multi-scale analysis of relationship between imperviousness and urban tree height using airborne remote sensing," *Remote Sens. Environ.*, vol. 194, pp. 391–400, Jun. 2017.

[2] R. Posselt, R. W. Mueller, R. Stöckli, and J. Trentmann, "Remote sensing of solar surface radiation for climate monitoring - the CM-SAF retrieval in international comparison," *Remote Sens. Environ.*, vol. 118, no. 6, pp. 186–198, Mar. 2012.

[3] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 116, pp. 24–41, Jun. 2016.

[4] L. Matikainen, K. Karila, J. Hyyppä, P. Litkey, E. Puttonen, and E. Ahokas, "Object-based analysis of multispectral airborne laser scanner data for land cover classification and map updating," *ISPRS J. Photogramm. Remote Sens.*, vol. 128, pp. 298–313, Jun. 2017.

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

[6] S. K. Pal, T. J. Majumdar, and A. K. Bhattacharya, "ERS-2 SAR and IRS-1C LISS III data fusion: A PCA approach to improve remote sensing based geological interpretation," *ISPRS J. Photogramm. Remote Sens.*, vol. 61, no. 5, pp. 281–297, Jan. 2007.

[7] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, "A fast intensity–hue–saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 309–312, Oct. 2004.

[8] 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, May 2010.

[9] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[10] Y. Kim, C. Lee, D. Han, Y. Kim, and Y. Kim, "Improved additive-wavelet image fusion," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 263–267, Mar. 2011.

[11] J. Ma and G. Plonka, "Computing with curvelets: From image processing to turbulent flows," *Comput. Sci. Eng.*, vol. 11, no. 2, pp. 72–80, Mar. 2009.

[12] B. Zhang, J. M. Fadili, and J. L. Starck, "Wavelets, ridgelets, and curvelets for poisson noise removal," *IEEE Trans. Image Process.*, vol. 17, no. 7, pp. 1093–1108, Jul. 2008.

[13] A. L. d. Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, Oct. 2006.

[14] X. Zhou, J. Liu, S. Liu, L. Cao, Q. Zhou, and H. Huang, "A GIHS-based spectral preservation fusion method for remote sensing images using edge restored spectral modulation," *ISPRS J. Photogramm. Remote Sens.*, vol. 88, no. 2, pp. 16–27, Feb. 2014.

[15] J. Liu, J. Huang, S. Liu, H. Li, Q. Zhou, and J. Liu, "Human visual system consistent quality assessment for remote sensing image fusion," *ISPRS J. Photogramm. Remote Sens.*, vol. 105, pp. 79–90, Jul. 2015.

[16] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled Shearlet transform," *Infrared Phys. Technol.*, vol. 73, pp. 286–297, Nov. 2015.

[17] H. Lin, Y. Tian, R. Pu, and L. Liang, "Remotely sensing image fusion based on wavelet transform and human vision system," *Int. J. Signal Process., Image Process. Pattern Recog.*, vol. 8, pp. 291–298, Jul. 2015.

[18] Z. Shao, J. Liu, and Q. Cheng, "Fusion of infrared and visible images based on focus measure operators in the curvelet domain," *Appl. Opt.*, vol. 51, no. 12, pp. 1910–1921, 2012.

[19] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.

[20] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse coding for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, Jun. 1996.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[21] M. Aharon, M. Elad, and A. Bruckstein, “K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation,” *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[22] B. Yang and S. Li, “Multifocus image fusion and restoration with sparse representation,” *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.

[23] B. Yang and S. Li, “Pixel-level image fusion with simultaneous orthogonal matching pursuit,” *Inf. Fusion*, vol. 13, no. 1, pp. 10–19, Jan. 2012.

[24] M. Ding, L. Wei, and B. Wang, “Research on fusion method for infrared and visible images via compressive sensing,” *Infrared Phys. Technol.*, vol. 57, no.2, pp. 56–67, Mar. 2013.

[25] 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.

[26] Q. Wei, J. Bioucas-Dias, and N. Dobigeon, and JY. 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.

[27] G. Pajares and J. M. de la Cruz, “A wavelet-based image fusion tutorial,” *Pattern Recog.*, vol. 37, no. 9, pp. 1855–1872, Sep. 2004.

[28] K. Fukushima, “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biol. Cybern.*, vol. 36, no. 4, pp. 193–202, Jan. 1980.

[29] K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Commun. ACM*, vol. 60, no. 6, pp. 84–90, Jun. 2017.

[31] W. Ouyang *et al.*, “Deepid-net: Deformable deep convolutional neural networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Boston, MA, USA, 2015, pp. 2403–2412.

[32] N. Zhang, J. Donahue, R. Girshick, and T. Darrel, “Part-based RCNNs for fine-grained category detection,” *presented at the European Conference Computer Vision*, Zurich, Switzerland, Sep. 2014.

[33] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 1988–1996.

[34] W. Ouyang and X. Wang, “Joint deep learning for pedestrian detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, 2013, pp. 2056–2063.

[35] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proc. 27th Int. Conf. Mach. Learn.*, Haifam, Israel, 2010, pp. 807–814.

[36] C. Dong, C. C. Loy, K. He, and X. Tang, “Image super-resolution using deep convolutional networks,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[37] J. Kim, J. K. Lee, and K. M. LEE, “Accurate image super-resolution using very deep convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Las Vegas, NV, USA, 2016, pp. 1646–1654.

[38] L. Wald, T. Ranchin, and M. Mangolini, “Fusion of satellite images of different spatial resolution: Assessing the quality of resulting images,” *Photogramm. Eng. Remote Sens.*, vol. 63, pp. 691–699, Nov. 1997.

[39] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, “Pansharpening by convolutional neural networks,” *Remote Sens*., vol. 8, no. 7, p. 594, Jul. 2016.

[40] F. Palsson, J. R. Seveinsson, and M. O. Ulfarsson, “Multispectral and hyperspectral image fusion using a 3D convolutional neural network,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 639–643, May 2017.

[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based leaning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[42] L. Alparone, L. Wald, J. Chanussot, C. Thomas, P. Gamba, and L. M. Bruce, “Comparison of pansharpening algorithms: outcome of the 2006 GRS-S datafusion contest,” *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3012–3021, Oct. 2007.

[43] L. Wald, “Quality of high resolution synthesised images: Is there a simple criterion?,” in *Proc. 3rd Conf. Fusion Earth Data: Merging Point Meas. Raster Maps Remotely Sensed Images*, Sophia-Antipolis, France, 2000, pp. 99–103.

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

[45] 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, pp. 193–200, Feb. 2008.

[46] H. Li, B. S. Manjunath, and S. K. Mitra, “Multi-sensor image fusion using the wavelet transform,” *Graph. Models Image Process.*, vol. 57, no. 3, pp. 235–245, May 1995.

[47] J. Cheng, H. Liu, T. Liu, F. Wang, and H. Li, “Remote sensing image fusion via wavelet transform and sparse representation,” *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 158–173, Jun. 2015.

[48] W. Dou, Y. Chen, X. Li, and D. Z. Sui, “A general framework for component substitution image fusion: An implementation using the fast image fusion method,” *Comput. Geosci.*, vol. 33, no. 2, pp. 219–228, Feb. 2007.

[49] Q. Liu, R. Hang, H. Song, and Z. Li, “Learning multiscale deep features for high-resolution satellite image scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 1, pp. 117–126, Jan. 2018.

[50] Y. Chen, H. Jiang, C. Li, X. Jia, and P. Ghamisi, “Deep feature extraction and classification of hyperspectral images based on convolutional neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 6232–6251, Oct. 2016.

[51] W. Li, G. Wu, F. Zhang, and Q. Du, “Hyperspectral image classification using deep pixel-pair features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 844–853, Feb. 2017.

[52] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, “Deep learning-based classification of hyperspectral data,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2014.

[53] F. P. S. Luss, B. P. Salmon, F. van den Bergh, and B. T. J. Maharaj, “Multiview deep learning for land-use classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2448–2452, Dec. 2015.

[54] S. Pierre, C. Soumith, and Y. Lecun, “Convolutional neural networks applied to house numbers digit classification,” in *Proc. Int. Conf. Pattern Recog.*, Tsukuba, Japan, 2012, pp. 3288–3291.

[55] J. Zhong, B. Yang, G. Huang, F. Zhong, and Z. Chen, “Remote sensing image fusion with convolutional neural network,” *Sens. Imag.*, vol. 17, no. 1, Dec. 2016, Art. no. 10.

[56] K. He, X. Zhang, S. Ren, and J. Sun, “Delving deep into rectifiers Surpassing human-level performance on imagenet classification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1026–1034.

[57] A. Garzelli, “Pansharpening of multispectral images based on nonlocal parameter optimization,” *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 2096–2107, Apr. 2015.

Zhenfeng Shao (M’15) received the Ph.D. degree in aerial photogrammetry from Wuhan University, China, in 2004.

He is currently a Professor in the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, China. His research interests are remote sensing and data mining.

Jiajun Cai received the B.Eng. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2016. He is currently working toward the M.Eng. degree in photogrammetry and remote sensing at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University.

His research interests include deep learning and image processing.