A Multiscale and Multidepth Convolutional Neural Network for Remote Sensing Imagery Pan-Sharpening

Qiangqiang Yuan, *Member, IEEE*, Yancong Wei, *Student Member, IEEE*, Xiangchao Meng, *Student Member, IEEE*, Huanfeng Shen, *Senior Member, IEEE*, and Liangpei Zhang, *Senior Member, IEEE*

Abstract—**Pan-sharpening is a fundamental and significant task in the field of remote sensing imagery processing, in which high-resolution spatial details from panchromatic images are employed to enhance the spatial resolution of multispectral (MS) images. As the transformation from low spatial resolution MS image to high-resolution MS image is complex and highly nonlinear, inspired by the powerful representation for nonlinear relationships of deep neural networks, we introduce multiscale feature extraction and residual learning into the basic convolutional neural network (CNN) architecture and propose the multiscale and multidepth CNN for the pan-sharpening of remote sensing imagery. Both the quantitative assessment results and the visual assessment confirm that the proposed network yields high-resolution MS images that are superior to the images produced by the compared state-of-the-art methods.**

Index Terms—**Convolutional neural network (CNN), multiscale feature learning, pan-sharpening, remote sensing.**

I. INTRODUCTION

IN REMOTE sensing images, panchromatic (PAN) images have a very high spatial resolution with the cost of lacking spectral band diversities. Multi-spectral (MS) images contain rich spectral information, but the levels of their resolution are times lower than that of PAN images. However, due to the technical limitations of sensors and other factors, remote sensing images with both high spatial and spectral resolutions, which are highly desirable in many remote sensing applications, are currently unavailable. Therefore, researchers have made efforts to fuse PAN images with MS images to produce an image with both high spatial and spectral resolutions, which is a process that is also called "pan-sharpening."

To date, a variety of pan-sharpening methods have been proposed, and most of them can be divided into three major categories:

A. Component Substitution (CS) Based Methods

This type of method traditionally transforms the MS image into a suitable domain. The specific component representing the spatial information of the MS image is then replaced by the PAN image, and inverse transformation is performed to reconstruct the fused image. Examples of CS-based methods are the typical intensity-hue-saturation fusion methods [1], [2], the principal component analysis fusion method [3], the Gram–Schmidt (GS) fusion method [4], and adaptive component-substitution-based satellite image fusion using partial replacement [5]. It should be noted that, in this group of methods, analysis of the correlation between the replaced MS component and the PAN image has a great influence on the fusion result.

B. Multiresolution Analysis (MRA) Based Methods

Compared with the traditional CS-based methods, the MRA-based methods generally have better spectral information preservation. In general, this type of method first extract the spatial structures from the PAN image by wavelet transform, Laplacian pyramid, etc., and then the extracted spatial structure information is injected into the up-sampled MS images to obtain the fused image. Examples of this type of method are the fusion methods based on wavelet transform [6] or curvelet transform [7], the analysis of modulation transfer function (MTF) [8], [9], and the smoothing filter based intensity modulation (SFIM) method [10]. A combination of CS and MRA has also been recently proposed to enhance the spatial-spectral unified fidelity of fused images [11]. However, these types of methods generally produce spatial distortion, and there is a strict requirement

Q. Yuan is with the School of Geodesy and Geomatics and the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: yqiang86@gmail.com).

Y. Wei is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: ycwei@whu.edu.cn).

X. Meng is with the Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China (e-mail: mengxiangchao@nbu.edu.cn).

H. Shen is with the School of Resource and Environmental Science and the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: shenhf@whu.edu.cn).

L. Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing and the Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

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

for accurate coregistration between the PAN and up-sampled MS images.

C. Optimization-Based (OB) Approaches

These types of methods are based on the image observation models and regard the solution of the fused image as an ill-posed inverse problem. Generally, the fusion images can be solved by minimizing a loss function with the prior constraints, such as the minimum mean square error based band-dependent spatial detail model [12], nonlocal optimization based on k-means clustering algorithm [13], Bayesian posterior probability [14], adaptive regularization based on normalized Gaussian distribution [15], total variation operators [16], [17], and sparse reconstruction based fusion methods [18]. Recently, another group of optimization based approaches using the advanced deep learning models are also proposed, which will be specifically introduced in the following.

Although a variety of pan-sharpening methods have been proposed, the disadvantages of these three major types of methods are hard to ignore. In the CS- and MRA-based fusion methods, the transformation from observed images to fusion targets is not rigorously modeled and distortion in the spectral domain is very common. In the results of the MBO-based methods, the spectral distortion can be reduced by better modeling of the transformation, and a much higher accuracy can be produced, but the linear simulation from the observed and fusion image is still a limitation, especially when the spectral coverages of the PAN and MS images do not fully overlap and lead to the fusion process being highly nonlinear. Furthermore, in the MBO-based methods, the design of the optimal fusion energy function is heavily reliant on prior knowledge, and on images with different distributions and quality degeneration, these models are not robust. Furthermore, solving the regularization models generally requires iterative computing, which is time-consuming and may cause incidental errors, especially for the images with a large size.

To overcome those shortcomings, advanced algorithms have been introduced in recent years, and among them, the deep learning models are some of the most promising approaches. Deep learning models are built with multiple transforming layers, and in each layer, its input is linearly filtered to produce an output, and multiple layers are stacked to form a total transformation with high nonlinearity. The most outstanding advantage of the deep learning models is that all the parameters included in the model can be updated under the supervision of training samples, and thus the requirement for prior knowledge is reduced and much higher fitting accuracies can be expected.

For both natural images and remote sensing images, in the field of most low-level vision tasks, e.g., image denoising, deblurring, superresolution, inpainting, etc. [23]–[31], deep learning based methods have achieved state-of-the-art accuracies in recent years, and their performances are continuously being improved. However, in the field of pan-sharpening, only limited studies have been undertaken in recent years to introduce deep learning models. Examples are the sparse deep neural network [32] and the pan-sharpening neural network (PNN) [33], the latter of which has achieved impressive performance gains.

However, as the design of the PNN is completely borrowed from the superresolution CNN (SRCNN) proposed in [22], which is considered a relatively simple and shallow architecture when compared with its later derivations [23], [27], [28], [30], there is still plenty of room for improvement. To exploit the advantages of deep learning and overcome the shortcomings of the current methods, we propose an original network that is specifically designed for the pan-sharpening task, while it can also be generalized for other types of image restoration problems. The framework consists of a PNN and a deeper multiscale neural network. The former network performs simple feature extraction, while the latter network contains multiscale feature extraction layers and builds a deep architecture. We believe that as the scale of features greatly varies among different ground objects from multiple sensors, introducing multiscale feature extraction can help to learn more robust convolutional filters, and thus the fusion accuracy can be advanced from the current state-of-the-art level. This assumption is fully supported by the experimental results, which are described in Section IV.

The rest of this paper is organized as follows. The background knowledge to pan-sharpening and the related deep learning works are introduced in Section II. The detailed architecture of the proposed multiscale and multidepth convolutional neural network (MSDCNN) is described in Section III. The results of the pan-sharpening accuracy assessment are presented in Section IV. Finally, a discussion and the conclusion are given in Section V.

II. Background

A. Pan-Sharpening Based on Linear Models

Assuming that the low-resolution MS image is considered as a degraded observation g_{MS}, then the PAN image g_{PAN} that matches g_{MS} is included to guide the prediction process of the high-resolution spatial details in the ground truth f_{MS}. The main aim of the pan-sharpening task is to preserve the unified spatial-spectral fidelity for the fused image. For a low-resolution MS image g_{MS} with S spectral bands, we denote the pan-sharpened result as F_{MS}, which is an estimation of f_{MS}, and then the constraint function of the MS image pan-sharpening can be formed as

$$\underset{F_{\mathrm{MS}}}{\arg\min} \sum_{i=1}^{S} \left\| f_{\mathrm{MS}(i)} - F_{\mathrm{MS}(i)} \right\|_2^2 \qquad (1)$$

where F_{MS} is obtained from a fusion function

$$F_{\mathrm{MS}} = P(g_{\mathrm{MS}}, g_{\mathrm{PAN}}) \qquad (2)$$

In (2), $P(.)$ represents the pan-sharpening process. In the traditional MBO approaches, both g_{MS} and g_{PAN} are considered as degraded observations of f_{MS} in relative domains, and the fusion process is simulated under a linear framework as

$$\begin{bmatrix} g_{\mathrm{MS}} \\ g_{\mathrm{PAN}} \end{bmatrix} = \begin{bmatrix} DH f_{\mathrm{MS}} \\ R f_{\mathrm{MS}} \end{bmatrix} + \begin{bmatrix} N_{\mathrm{MS}} \\ N_{\mathrm{PAN}} \end{bmatrix} \qquad (3)$$

where D is a down-sampling matrix in the spatial domain, and similarly, R is the spectral response matrix of the PAN

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

YUAN *et al.*: MULTISCALE AND MULTIDEPTH CONVOLUTIONAL NEURAL NETWORK FOR REMOTE SENSING IMAGERY PAN-SHARPENING 3
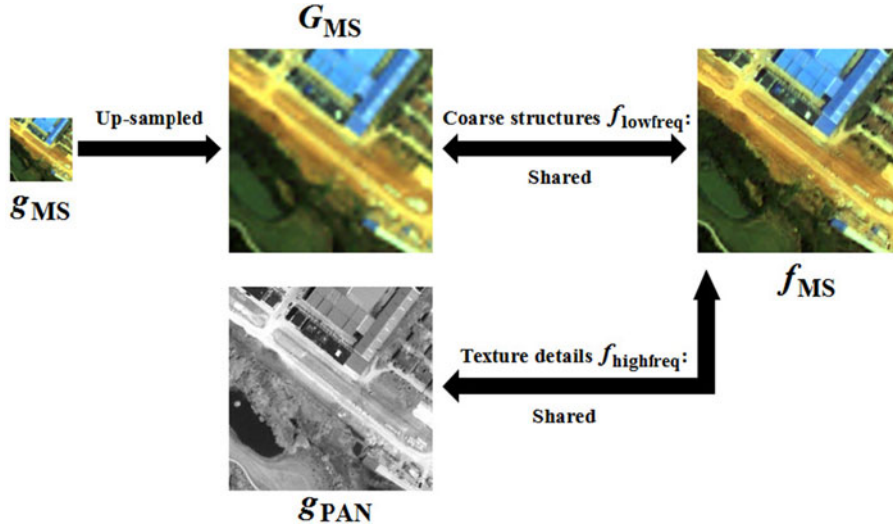

Fig. 1. Visual correlation between a low-resolution MS image, a PAN image, and a high-resolution MS image.

channel of the sensor, which down-samples the latent ground truth along the spectrum. H is a blurring matrix, while N_{MS} and N_{PAN} are the additive noise, which is assumed to be Gaussian distributed. Therefore, (2) is linearly fitted by solving an optimization function as

$$\underset{F_{MS}}{\arg\min}\{\lambda_1 \left\|DHF_{MS} - g_{MS}\right\|_P^2 + \lambda_2 \left\|RF_{MS} - g_{PAN}\right\|_P^2$$
$$+ \lambda_3 \varphi(F_{MS})\} \qquad (4)$$

in which $\lambda_i (i = 1, 2, 3)$ represents the weights that control the contributions of the fidelity items and the constraint operator $\varphi(F_{MS})$, the latter of which is based on reasonable assumptions and prior knowledge to reduce the ill-posed property of the problem.

However, it should be noted that in the pan-sharpening process, the bandwidths of the PAN and MS images are not guaranteed to fully overlap. For example, the MS bandwidth of WorldView-2 ranges from 400 to 1040 nm and is divided into eight bands, and its PAN bandwidth covers 450–800 nm. Thus, if we keep simulating the transformation $P(.)$ from a linear perspective, as in (4), it is difficult to merge the down-sampled spectra of the PAN images into the spectra of the MS images while preserving the fidelity of the latter. The drawbacks of such linear models can be explained as follows. First, a satisfactory accuracy can rarely be achieved when linear functions are employed to fit complex transformations, especially for ill-posed inverse problems. Second, prior knowledge that has been artificially introduced into the problem, e.g., the design of $\varphi(F_{MS})$, is not guaranteed to be suitable for general tasks and may increase the system error. Furthermore, for images of many complex circumstances and from different sensors, the value of λ_i needs to be empirically chosen and lacks a robust solution. Thus, the abilities of the linear optimization models are somewhat limited.

To overcome the drawbacks of the linear models, a nonlinear function is needed to fit the fusion process, which requires us to employ a different point of view to investigate the correlation among g_{MS}, g_{PAN}, and f_{MS}. Therefore, the idea of deep learning is adopted, and is introduced in the next subsection.

B. Deep Learning for Pan-Sharpening

As illustrated in Fig. 1, for the texture details contained in g_{PAN}, we regard them as high-frequency components of f_{MS}, and the coarse spatial structures of G_{MS} are regarded as low-frequency components. Thus, we can employ a filtering function to extract the features $f_{lowfreq}$ and $f_{highfreq}$, and merge them to yield the high-resolution estimation F_{MS}.

How do we obtain a set of filters that can accurately extract complex features from various ground scenes, without causing spectral distortion? The recently developed deep learning approach is one of the most advanced answers to this problem. In the different deep learning networks, convolutional neural networks (CNNs) are a branch of the deep learning models that has impressively swept the field of computer vision and image processing in recent years. In this paper, it is introduced as a prototype of our proposed methodology. Compared with the traditional hand-crafted extractors for features, the superiority of CNNs can be explained with two concepts—"deep" and "learning"—which are explained in the following.

Deep: The architectures of CNNs are formed by stacking multiple convolutional layers. Although each of these layers functions as a linear filtering process, a whole network is able to fit a very complex nonlinear transformation that maps $\{G_{MS}, g_{PAN}\}$ to f_{MS}. The nonlinearity and fitting ability of CNNs are not limited to a certain level, as the depth of the network can be infinitely expanded along the direction in which the layers are stacked.

Learning: To extract features from G_{MS} and g_{PAN}, the filtering process in every convolutional layer of a CNN is executed using convolutional kernels. With the supervision of f_{MS} as a target, the network iteratively updates all the kernels to seek an optimal allocation, and thus it is defined as a "learning" process. When the loss between f_{MS} and F_{MS} reaches a satisfactory

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

4

IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING
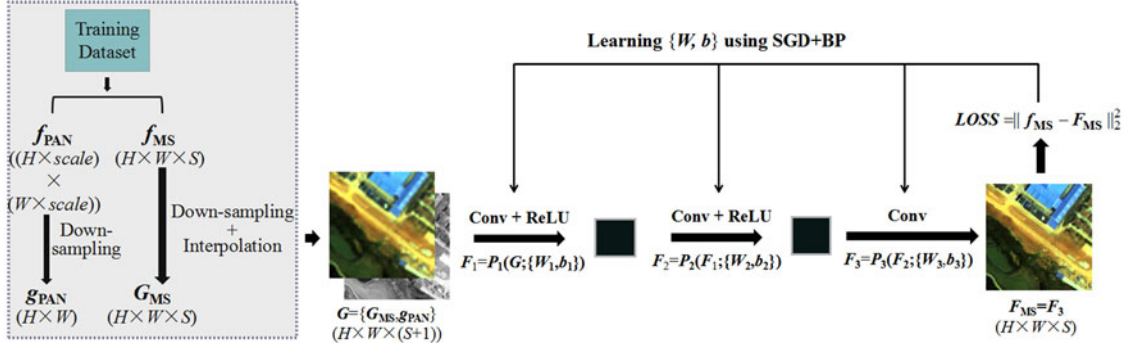

Fig. 2. Flowchart of basic CNN-based pan-sharpening.
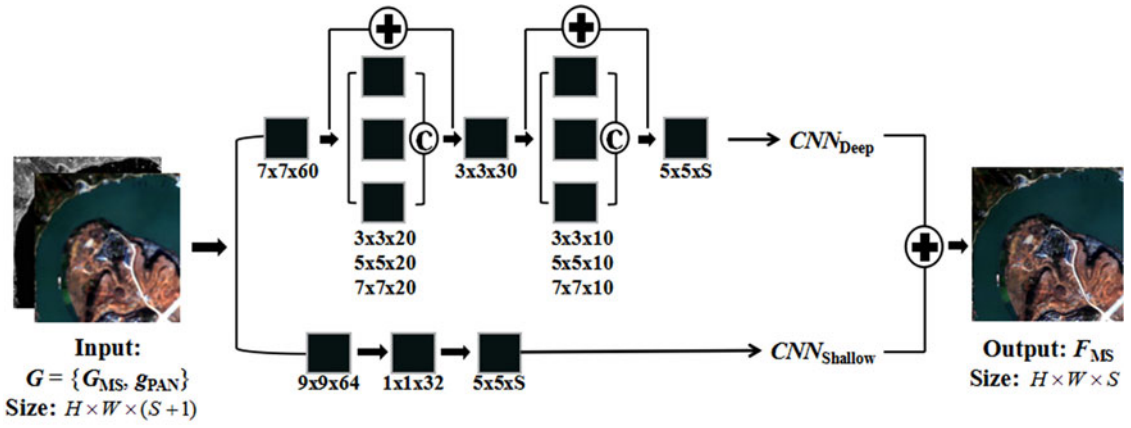

Fig. 3. Flowchart of passing the input MS image through MSDCNN to yield a fused result.

convergence, the learning of the network is finished and an accurate end-to-end function is obtained for the pan-sharpening. The flowchart of training a deep CNN on a training dataset is shown in Fig. 2.

Pan-sharpening with a basic CNN: As mentioned above, G_{MS} and g_{PAN} are fed into a CNN to directly yield a fused image F_{MS}. In the network, the input images are passed through L layers, and the filtering process executed in the nth layer can be described as

$$F_n = P_n(F_{n-1}) \qquad (5)$$

where F_n is the output of the nth layer. Thus, the fusion process can be described as follows:

$$F_0 = G = \{G_{MS}, g_{PAN}\}, \text{Size} : H \times W \times (S+1) \quad (6)$$

$$F_n = P_n(F_{n-1}) = ReLU(W_n \circ F_{n-1} + b_n),$$
$$\text{Size} : H \times W \times C_n, n = 1, \dots, L-1 \qquad (7)$$

$$F_{MS} = F_L = W_L \circ F + b_L, \text{Size} : H \times W \times S \quad (8)$$

where $\circ$ represents three-dimensional convolution, which is the feature extractor in $P_n(F_{n-1})$, and W_n contains C_n groups of convolutional kernels, where the size of each group is $h_n \times w_n \times C_{n-1}$, and b_n is a bias vector with the size of $1 \times 1 \times C_n$. Thus, for the nth layer, C_n represents the spectral dimensionality of its output and can be artificially set. The rectified linear unit

(ReLU) is used to introduce nonlinearity in the function

$$ReLU(x) = \max(x, 0). \qquad (9)$$

III. Proposed Network: MSDCNN

Based on the basic architecture of a CNN with three convolutional layers for pan-sharpening, as previously mentioned, we introduce two concepts to improve the architecture of the network: the *multi-scale feature extraction block* and *skip connection*. The proposed MSDCNN contains two subnetworks: A fundamental three-layer CNN with the same architecture as in [22] and [33], and a deeper CNN with two multiscale convolutional layer blocks. The whole architecture of MSDCNN is displayed in Fig. 3.

A. Multiscale Feature Extraction Block

As mentioned before, the coarse structures and texture details are the features that need to be extracted from ground objects and scenes. In remote sensing imagery with a meter- or submeter-level spatial resolution, the sizes of the ground objects vary from very small neighborhoods to large regions containing thousands of pixels, and a ground scene may cover many objects with various sizes. From the feature maps displayed in Fig. 4, it is indicated that the features with a smaller scale, such as the short edges of buildings and the textures of vegetation, tend to

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

YUAN *et al.*: MULTISCALE AND MULTIDEPTH CONVOLUTIONAL NEURAL NETWORK FOR REMOTE SENSING IMAGERY PAN-SHARPENING 5
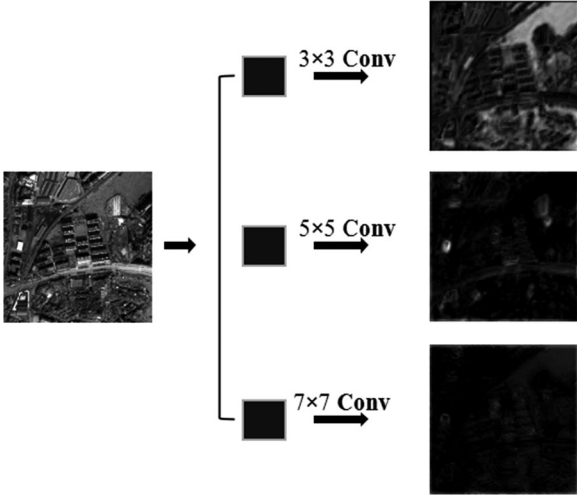

Fig. 4. Feature maps extracted by convolutional filters with three different sizes, which are selected from the first layer of a trained MSDCNN model.
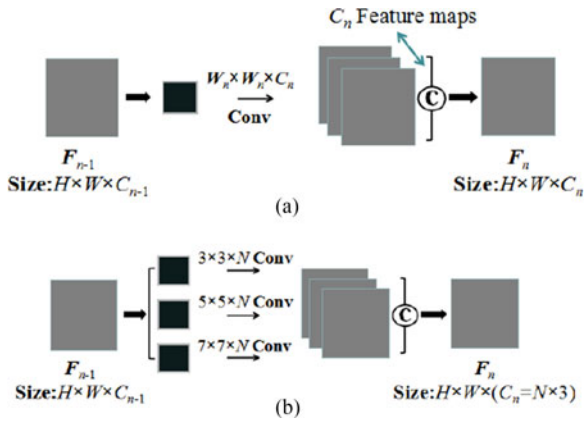

Fig. 5. Difference between a basic convolutional layer and a layer for multiscale feature extraction, where $\mathbf{C}$ stands for concatenating images along the spectral dimension. (a) Basic convolutional layer. (b) Convolutional layer for multiscale feature extraction.

Fig. 6. Complete architecture of the proposed multiscale convolutional layer block with a short-distance skip connection.

respond to convolutional filters with a smaller size, while the coarse structures tend to be extracted by larger filters.

To make adequate use of the rich spatial information in high-resolution imagery and improve the robustness of the feature extraction among various and complex ground scenes, we introduce the multiscale convolutional layer block, which was applied to image superresolution in [30] and classification in [35].

As illustrated in Fig. 5, in the nth layer, three sizes are set for the convolutional kernels contained in the multiscale layer block: $3 \times 3, 5 \times 5$ and 7×7. For each size, N groups of kernels are employed to produce N feature maps, and they are concatenated along the spectral dimension to form the output.

B. Skip Connection

As discussed in Section II-B, in CNNs, stacking more layers can lead to higher nonlinearity and can help to fit complex transformations more accurately. Visualized feature maps show that when an image is passed through a deeper network, the features extracted from it can b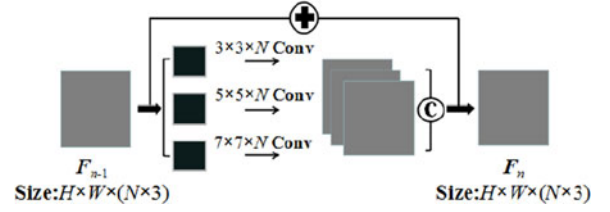e more abstract and representative [36], [37]. However, there is a significant problem in that in the training process of a deep CNN, the gradients of the loss to the network parameters are severely diminished during the back-propagation from output to input. Thus, in layers that are close to the input, updating of the convolutional kernels and bias vectors becomes too slow to reach the optimal allocation of all parameters.

In [22] and [33], it was indicated that for the fundamental architecture of a CNN, $L = 3$ is an upper limit to the depth of the network, and adding more layers can no longer boost the accuracy performance, while the increase in training time also becomes unacceptable. To deal with the problem, residual learning [38] is now considered to be one of the most effective solutions for training deep CNNs, in which the convolutional filtering process $\boldsymbol{F}_n = \boldsymbol{P}_n(\boldsymbol{F}_{n-1})$ is replaced with $\boldsymbol{F}_n = \boldsymbol{F}_{n-1} + \boldsymbol{P}_n(\boldsymbol{F}_{n-1})$, and thus the residual $\boldsymbol{F}_n - \boldsymbol{F}_{n-1}$ becomes the target of the prediction. This simple and effective architecture is called a "skip connection." It is assumed that the distribution of features in the residual image is very sparse and most of the pixel values are close to zero. Thus, the loss-parameters surface of a residual learning function becomes much smoother than the surface of a regular CNN, and the distances from the local minimum points to the optical minimum are shortened.

In [27], an end-to-end skip connection $\boldsymbol{F} = \boldsymbol{G} + \boldsymbol{P}(\boldsymbol{G})$ was designed to train a very deep CNN for image superresolution, aiming to use the whole network to directly predict the residual image $\boldsymbol{f} - \boldsymbol{G}$ from the input low-resolution image $\boldsymbol{G}$. However, for the pan-sharpening task, the end-to-end architecture is not suitable due to the different sizes of $\boldsymbol{G} = \{\boldsymbol{G}_{\mathbf{MS}}, \boldsymbol{g}_{\mathbf{PAN}}\}$ (size: $H \times W \times (S+1)$) and $\boldsymbol{f}_{\mathbf{MS}}$ (size: $H \times W \times S$). Thus, in the proposed network, a connection that only skips one layer is set for the block, as illustrated in Fig. 6.

C. Joint Learning for MSDCNN

As described in Fig. 3, the images output from the two subnetworks of MSDCNN are summed for a final estimation

$$
\begin{aligned}
\boldsymbol{F}_{\mathbf{MS}} &= \mathbf{MSD}(\boldsymbol{G}, \{\boldsymbol{W}, \boldsymbol{b}\}) \\
&= \mathbf{CNN}_{\mathbf{shallow}}(\boldsymbol{G}; \{\boldsymbol{W}_{\mathbf{shallow}}, \boldsymbol{b}_{\mathbf{shallow}}\}) \\
&+ \mathbf{CNN}_{\mathbf{deep}}(\boldsymbol{G}; \{\boldsymbol{W}_{\mathbf{deep}}, \boldsymbol{b}_{\mathbf{deep}}\}) \quad (10)
\end{aligned}
$$

where all the parameters contained in MSDCNN are jointly learned

$$
\underset{\boldsymbol{W}, \boldsymbol{b}}{\arg \min} \sum_{i=1}^{S} \left\| \boldsymbol{f}_{\mathbf{MS}(i)} - (\mathbf{MSD}(\boldsymbol{G}, \boldsymbol{W}, \boldsymbol{b}))_{(i)} \right\|_2^2. \quad (11)
$$

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

6   IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING

TABLE I
DETAILS OF THE THREE DATASETS USED IN THE TRAINING AND TESTING

Sensor	MS bands	Scenes	Covered regions	Training	Simulated experiments	Real-data experiments
QuickBird	4	4	Nanchang, China (for training) Shenzhen, China (for training) Wuhan, China (for testing) Yichang, China (for testing)	Patches: 51648 Input size: $41 \times 41 \times 5$ Output size: $41 \times 41 \times 4$	Patches: 160 Input size: $250 \times 250 \times 5$ Output size: $250 \times 250 \times 4$	Not included
IKONOS	4	7	Wuhan, China (all for testing)	Not included	Not included	Patches: 112 Input size: $400 \times 400 \times 5$ Output size: $400 \times 400 \times 4$
WorldView-2	8	4	San Francisco, United States (two scenes for training, two scenes for testing)	Patches: 59840 Input size: $41 \times 41 \times 9$ Output size: $41 \times 41 \times 8$	Patches: 80 Input size: $250 \times 250 \times 9$ Output size: $250 \times 250 \times 8$	Patches: 28 Input size: $800 \times 800 \times 9$ Output size: $800 \times 800 \times 8$

To iteratively learn the optimal allocation of $\{W, b\}$, we let $\{W^t, b^t\}$ represent the values of $\{W, b\}$ in the tth iteration after random initialization, and F_{MS}^t stands for the output from $\{W^t, b^t\}$. The current loss is then

$$\mathbf{LOSS}^t = \sum_{i=1}^{S} \left\| \boldsymbol{f}_{\mathrm{MS}(i)} - \boldsymbol{F}_{\mathrm{MS}(i)}^t \right\|_2^2 \qquad (12)$$

By computing the derivatives of $\mathbf{LOSS}^t$ to $\{W^t, b^t\}$, the gradients are obtained as

$$\left\{ \delta W^t, \delta b^t \right\} = \left\{ \begin{array}{l} \left. \frac{\partial \mathbf{LOSS}^t(\{W,b\};G)}{\partial W} \right|_{W=W^t, b=b^t} \\ \left. \frac{\partial \mathbf{LOSS}^t(\{W,b\};G)}{\partial b} \right|_{W=W^t, b=b^t} \end{array} \right\}. \qquad (13)$$

Stochastic gradient descent (SGD) is also applied as an effective way to accelerate the training process. Instead of computing the gradient for a single image, a batch of input images $\{G_1, \ldots, G_{\mathrm{Batchsize}}\}$ are fed into the network in the tth iteration to yield multiple outputs $\{F_{\mathrm{MS}_1}^t, \ldots, F_{\mathrm{MS}_{\mathrm{Batchsize}}}^t\}$ and an average loss is defined as

$$\overline{\mathbf{LOSS}}^t = \frac{1}{\mathrm{Batchsize}} \sum_{b=1}^{\mathrm{Batchsize}} \sum_{i=1}^{S} \left\| \boldsymbol{f}_{\mathrm{MS}_b(i)} - \boldsymbol{F}_{\mathrm{MS}_b(i)}^t \right\|_2^2. \qquad (14)$$

An input image is then randomly picked from $\{G_1, \ldots, G_{\mathrm{Batchsize}}\}$ and used as G in (10) for computing the gradients. With $\{\delta W^t, \delta b^t\}$ known, $\{W^t, b^t\}$ can be updated using a classic momentum (CM) algorithm [39]. We let $\boldsymbol{\theta} = \{W, b\}$ represent all the parameters in the network, and then the updating of $\boldsymbol{\theta}$ as follows:

$$\Delta \boldsymbol{\theta}^t = \mu \cdot \Delta \boldsymbol{\theta}^{t-1} - \varepsilon \cdot \delta \boldsymbol{\theta}^t \qquad (15)$$

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \Delta \boldsymbol{\theta}^t \qquad (16)$$

where μ is the momentum and ε is the learning rate. During the training process, gradient clipping is also necessary to avoid gradient explosion. In each iteration, a summed L2-norm of all the gradients is limited, which means that δW^t and δb^t are clipped as

$$\left\{ \delta W^t, \delta b^t \right\}_{\mathrm{Clipped}} = \left\{ \frac{\delta W^t}{\|\delta W^t\|_2^2 / 0.1}, \frac{\delta b^t}{\|\delta b^t\|_2^2 / 0.1} \right\}. \qquad (17)$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Settings

1) Datasets: To simulate the fusion transformation, original MS images with different numbers of spectral bands from QuickBird and WorldView-2 sensors were used as the ground truth $\boldsymbol{f}_{\mathrm{MS}}$, and we then down-sampled $\boldsymbol{f}_{\mathrm{MS}}$ and used bicubic interpolation to obtain the low-resolution MS image G_{MS}. The PAN image was also down-sampled as g_{PAN}, and thus the ratios of the scales among G_{MS}, g_{PAN}, and $\boldsymbol{f}_{\mathrm{MS}}$ were kept the same to the real situation.

For training and simulated testing of the proposed MSD-CNN, we collected two large datasets from QuickBird and WorldView-2 images, which were divided into smaller patches to separately train two networks with different numbers of input bands. Details of the datasets used in the experiments are listed in Table I. It should be noted that the number of quantitatively tested samples included in our datasets (two datasets for the quantitative assessment, 240 images in total, with a spatial size of 250×250) was much larger than in the referenced papers; for instance, in [19], three datasets and three images with a spatial size of 600×600 were used, and in [33], three datasets and 150 images with a spatial size of 320×320 were considered.

For the real-data experiments, another smaller dataset was collected from a group of IKONOS images to test the network, and the network was tested on the WorldView-2 dataset with eight bands. The 112 patches in the real-data experiment for images with four bands were collected by fully segmenting 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.

YUAN *et al.*: MULTISCALE AND MULTIDEPTH CONVOLUTIONAL NEURAL NETWORK FOR REMOTE SENSING IMAGERY PAN-SHARPENING 7

TABLE II
NUMERIC ASSESSMENT OF THE SIMULATED QUICKBIRD IMAGE
PAN-SHARPENING

Bands	Algorithm	PSNR ($\uparrow$)	Q ($\uparrow$)	ERGAS ($\downarrow$)	SAM ($\downarrow$)	Q4 ($\uparrow$)
4	GS [4]	34.0907	0.8305	4.5014	4.0227	0.6831
	PRACS [5]	35.9282	0.8397	3.7501	3.5646	0.6138
	MTF-GLP [8]	34.3894	0.8227	4.4409	3.7893	0.6803
	SFIM [10]	34.4410	0.8264	5.1491	3.7708	0.6818
	AWLP [42]	34.2055	0.8314	4.0463	3.6587	0.6466
	TSSC [19]	35.3860	0.8488	3.9773	3.7154	0.7039
	PNN [33]	38.5201	0.9206	2.7110	2.6405	0.7569
	MSDCNN	**39.2674**	**0.9303**	**2.5408**	**2.4605**	**0.7924**

TABLE III
NUMERIC ASSESSMENT OF THE SIMULATED WORLDVIEW-2 IMAGE
PAN-SHARPENING

Bands	Algorithm	PSNR ($\uparrow$)	Q ($\uparrow$)	ERGAS ($\downarrow$)	SAM ($\downarrow$)	Q8 ($\uparrow$)
8	GS [4]	33.6506	0.8606	4.8395	6.1412	0.5781
	PRACS [5]	35.7979	0.8631	4.5579	6.2920	0.6849
	MTF-GLP [8]	34.8187	0.8788	4.3748	5.7698	0.6324
	SFIM [10]	34.8078	0.8756	4.3230	5.7579	0.6284
	AWLP [42]	35.0906	0.8769	4.4214	5.9263	0.6870
	TSSC [19]	36.7291	0.8951	3.9735	5.8269	0.6941
	PNN [33]	37.7634	0.9389	3.0695	4.4757	0.7697
	MSDCNN	**38.1045**	**0.9570**	**2.9331**	**4.2483**	**0.7740**

seven scenes of IKONOS images, while the 28 patches in the real-data experiment for images with eight bands were selected from the two test scenes of WorldView-2 images, covering regions of impervious surfaces, water bodies, and urban vegetation.

2) Model Implementation: For each dataset, MSDCNN was trained for 300 epochs (about 250 000 iterations), and the batch size was set to 64. To apply CM with SGD, $\mu = 0.9$ and $\varepsilon = 0.1$ were used as the default settings. With the Caffe [40] deep learning framework supported by a GPU (NVIDIA Quadro M4000) and CUDA 7.5, the training process for each model cost roughly 8 h.

Testing of all the convolutional networks was performed with the support of MatConvNet [41] on a Dell Tower 7810 workstation with an Intel CPU (Xeon E5-2620 v3 @ 2.40 GHz).

3) Compared Algorithms: For the numeric and visual assessment, seven traditional and state-of-the-art algorithms were used, representing different branches of pan-sharpening methods: GS [4] and partial replacement adaptive component substitution (PRACS) [5] belonging to CS; the MTF-based generalized Laplacian pyramid (MTF-GLP) [8], SFIM [10], and additive wavelet luminance proportion (AWLP) [42] belonging to multiresolution analysis; two-step sparse coding (TSSC) [19] based on regularization constraint model; and in the deep learning field, the PNN [33] based on a basic CNN containing three layers was considered as the main competitor to the proposed MSDCNN. We are thankful to Vivone *et al.* [43] for providing the toolbox that helped us to implement five of the seven referenced algorithms, except TSSC and PNN.

B. Simulated Experiments

In these experiments, the PAN and MS images were downsampled to simulate the low-resolution input g_{MS} and g_{PAN}, while the original MS images were employed as the ground truth f_{MS} to assess the qualities of the pan-sharpened results. Five numeric metrics were applied to quantify the qualities of the pan-sharpened images from the simulated experiments: the peak signal-to-noise ratio (PSNR) [44], the universal image quality metric (Q) [45], the *Erreur Relative Globale Adimensionnelle de Synthèse* (ERGAS) [46], the spectral angle mapper (SAM) [47], and $Q2^n$: An expanded version of Q that adds spectral fidelity into consideration [48]. The results of the simulated experiments are listed in Tables II and III, and in each comparison group, the best performance is marked in bold.

From the numeric assessment results listed previously, the superiority of the two CNN-based algorithms compared with the traditional methods is clear, as under all the full-reference metrics, the performances of PNN and MSDCNN are far ahead of the other algorithms, while the lead status is held by MSDCNN. For the 240 tested image patches containing various ground objects, the impressive performance gains of the proposed network helps us to confirm that the multiscale convolutional layer blocks significantly contribute to improving the robustness of the feature extraction and merging in all the bands along the spectral dimension.

As numeric metrics are applied to assess the quality of fused images from a quantifiable perspective, careful visual inspection is also needed to identify artifacts and distortions that elude the quantitative analysis. From the results of the simulated experiments, two groups of images that typically highlight the advantages and drawbacks of the various methods are selected and displayed in Figs. 7–8. For the purpose of displaying true-color images, the spectral bands covering the wavelengths of red, blue, and green light are selected according to the MS band division of the sensor, i.e., the 3rd, 2nd, and 1st bands of Quick-Bird, and the 5th, 3rd, and 2nd bands of WorldView-2.

By comparing the images displayed in Figs. 7 and 8, it can be seen that the results of the CNN-based methods are the most similar to the ground truth, both in spatial detail and spectral fidelity. For example, the vegetation areas in the lower-right of the group of images listed in Fig. 7. In particular, the proposed MSDCNN performs better than PNN [33] in preserving edges and the spectral features of ground objects with very small sizes, such as the concrete area in the middle-left of Fig. 7(h)–(i) and the bare soil in the upper-middle of Fig. 8(h)–(i). In some of the other six methods, while the spatial details are impressively sharpened and highlighted, noticeable spectral distortion is also apparent (GS [4], AWLP [42], MTF-GLP [8], and SFIM [10]). In contrast, better colors are obtained in the results of PRACS [5], but the restoration of spatial information is still not satisfactory. Fig. 7(e) shows that for the QuickBird dataset, TSSC [19] is a well-balanced solution, but when it came to the WorldView-2 dataset, there is still a gap between the performance of the sparse representation based model and the proposed MSDCNN.

The comparisons strongly support our statement that for remote sensing images with multiple sources that do not fully overlap in the spectral domain, nonlinear models based on deep learning are better able to handle the fusion task. It should also

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

8   IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING
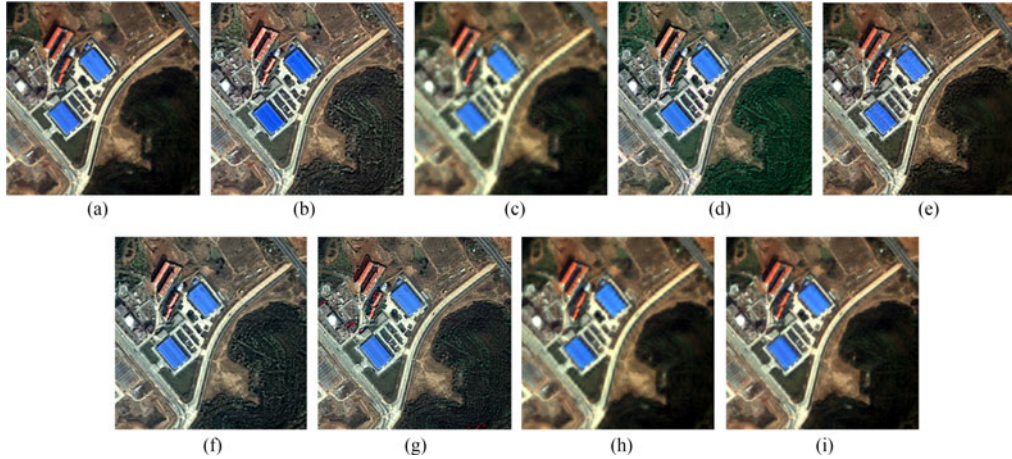

Fig. 7. Results of the simulated experiment on an area of industrial land, which was extracted from a QuickBird image of Yichang, China, obtained in 2015. (a) Ground truth. (b) GS [4]. (c) PRACS [5]. (d) AWLP [42]. (e) TSSC [19]. (f) MTF-GLP [8]. (g) SFIM [10]. (h) PNN [33]. (i) MSDCNN.
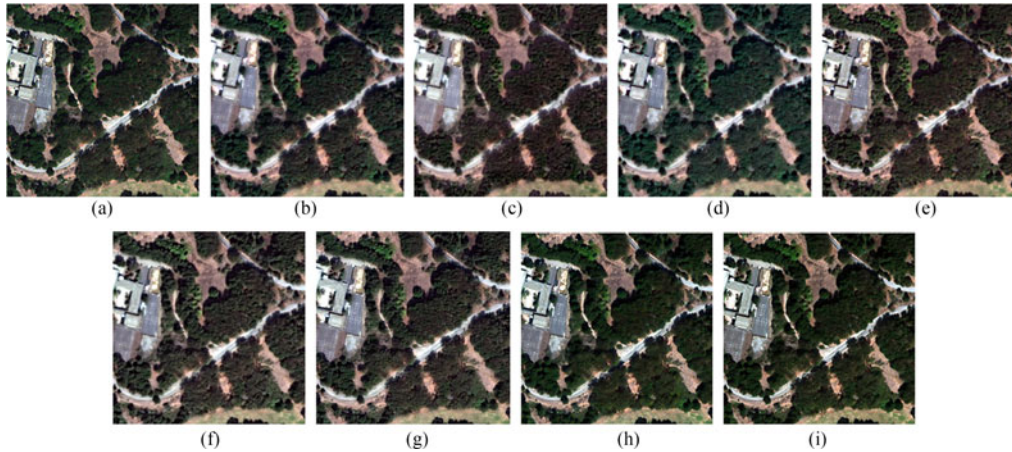

Fig. 8. Results of the simulated experiment on an area of city vegetation, which was extracted from a WorldView-2 image of San Francisco, United States, obtained in 2011. (a) Ground truth. (b) GS [4]. (c) PRACS [5]. (d) AWLP [42]. (e) TSSC [19]. (f) MTF-GLP [8]. (g) SFIM [10]. (h) PNN [33]. (i) MSDCNN.

TABLE IV
NUMERIC ASSESSMENT OF REAL-DATA IKONOS AND WORLDVIEW-2 IMAGE PAN-SHARPENING

IKONOS

Bands	Algorithm	QNR ($\uparrow$)	D_S ($\downarrow$)	D_λ ($\downarrow$)
4	GS [4]	0.7661	0.1753	0.0729
	PRACS [5]	0.8451	0.1183	0.0445
	MTF-GLP [8]	0.7434	0.1580	0.1202
	SFIM [10]	0.7526	0.1601	0.1068
	AWLP [42]	0.7433	0.1634	0.1148
	TSSC [19]	0.8587	0.0997	0.0497
	PNN [33]	0.8606	0.0895	0.0555
	MSDCNN	**0.8797**	**0.0774**	**0.0469**

WorldView-2

Bands	Algorithm	QNR ($\uparrow$)	D_S ($\downarrow$)	D_λ ($\downarrow$)
8	GS [4]	0.8403	0.1264	0.0415
	PRACS [5]	**0.8916**	0.0892	**0.0224**
	MTF-GLP [8]	0.8208	0.1108	0.0797
	SFIM [10]	0.8380	0.1073	0.0645
	AWLP [42]	0.8458	0.0991	0.0635
	TSSC [19]	0.8425	0.1037	0.0617
	PNN [33]	0.8725	0.0826	0.0538
	MSDCNN	0.8893	**0.0779**	0.0390

be noted that compared with the related PAN image and some of the over-sharpened fusion results, the slightly "blurry" appearance is also shared by the ground truth and the result of MSDCNN, which indicates that instead of being constrained by artificially given priors, the proposed network is able to fit various types of transformation.

C. Real-Data Experiments

Original MS and PAN images were also input into the models to yield full-resolution results. There are nonreference numeric metrics that can quantify the qualities of pan-sharpened images, i.e., the quality with no-reference index (QNR) [49] and the spatial and spectral components of it (D_S and D_λ). We employed the three metrics for the quantitative assessment of the real-data experiments, and the results are listed Table IV.

However, considering that these metrics are computed with G_{MS} and g_{PAN} as references, instead of the unattainable ground truth, we should note that what can be quantified by such metrics is the similarity of certain components in the fused images to the low-resolution observations, but not the real fidelity at the level of high resolution. The comparisons in

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

YUAN *et al.*: MULTISCALE AND MULTIDEPTH CONVOLUTIONAL NEURAL NETWORK FOR REMOTE SENSING IMAGERY PAN-SHARPENING   9
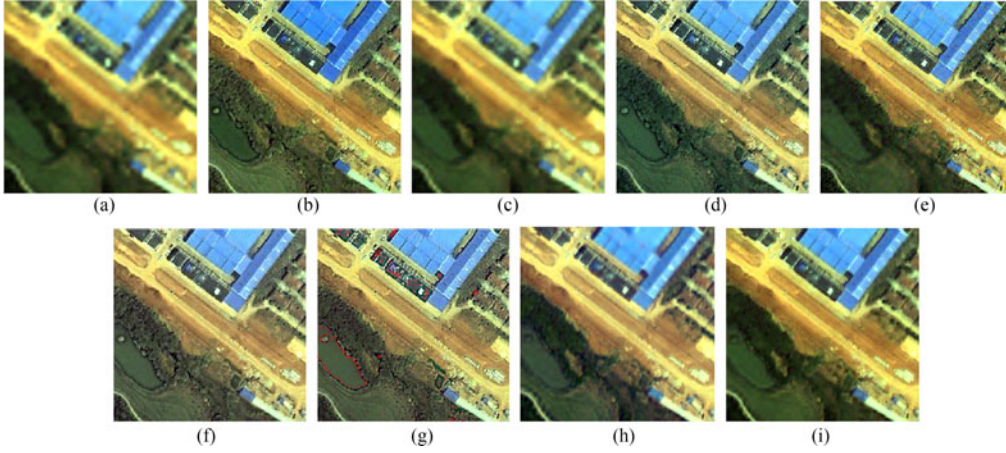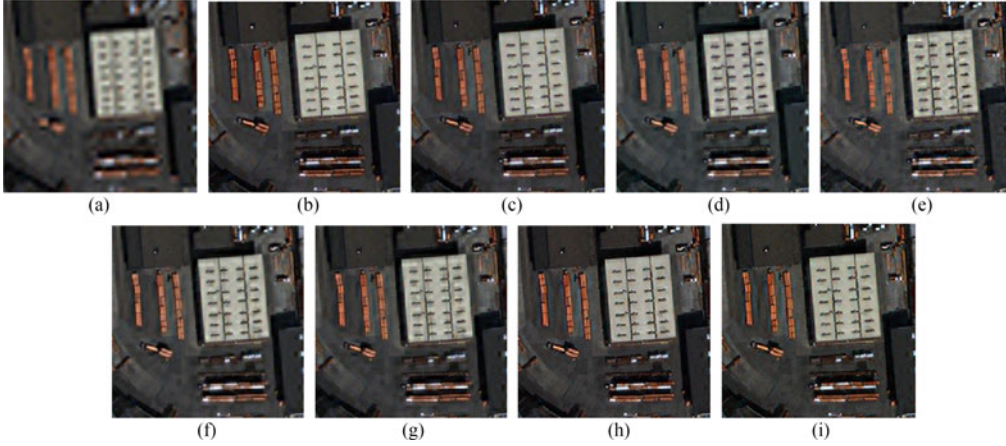
Fig. 9. Results of the real-data experiment on an area of industrial land, which was extracted from an IKONOS image of Wuhan, China. (a) Bicubic. (b) GS [4]. (c) PRACS [5]. (d) AWLP [42]. (e) TSSC [19]. (f) MTF-GLP [8]. (g) SFIM [10]. (h) PNN [33]. (i) MSDCNN.

Fig. 10. Results of the real-data experiment on an area of impervious surface, which was extracted from a WorldView-2 image of San Francisco, United States, obtained in 2011. (a) Bicubic. (b) GS [4]. (c) PRACS [5]. (d) AWLP [42]. (e) TSSC [19]. (f) MTF-GLP [8]. (g) SFIM [10]. (h) PNN [33]. (i) MSDCNN.

Table IV also support our assumption, as the results of PRACS [5] are very similar to the related low-resolution MS images and barely sharpened in the spatial domain, but by the similarity, they achieved very high D_λ values and jointly improved their QNR index to a state-of-the-art level.

Thus, in the following discussion, the real-data experiments are mainly discussed based on the visual inspection, instead of the three numeric metrics. Three ground regions were selected from the pan-sharpened full-resolution images to be investigated, as displayed in Figs. 9–11.

By comparing the images displayed in Fig. 9, we can observe a tendency that is similar to the story told by the previous simulated experiments: MSDCNN and PNN [33] return images with the best spectral fidelity and appropriately sharpened spatial details, while the proposed network performs slightly better in preserving details with small sizes. Among the other compared methods, TSSC [19] remains competitive in the real-data experiments, which is supported by the high quality of Fig. 9(e) and its high similarity to the related image obtained by MSDCNN in Fig. 9(i). However, when it comes to the WorldView-2 dataset, as shown in Figs. 10(e) and 11(e), the performance of TSSC becomes less robust, while MSDCNN is still able to avoid introducing ringing artifacts from the up-sampled MS images and

prevents spectral distortion, for example, the impressive quality of Fig. 11(i) shows that, though the MS image in Fig. 11(a) is severely corrupted after interpolation, our proposed network still performed a good fusion with the guidance from its related PAN image.

D. Further Discussion

In this subsection, the default settings of MSDCNN used in the experiments are compared with the alternatives. The performance of the network with different settings was tested by simulated experiments on the QuickBird dataset containing 160 images and assessed with the full-reference Q and ERGAS metrics.

1) Setting Hyper-Parameters for Training MSDCNN: As mentioned above, the momentum and learning rate are initialized as $\mu = 0.9$ and $\varepsilon = 0.1$, and for every 60 epochs, ε is multiplied by $\gamma = 0.5$, while μ is fixed as 0.9. From the performance-to-epoch curves in Fig. 12, we can see that the residual learning architecture of MSDCNN helps the network to quickly reach state-of-the-art accuracy within about 50 training epochs, while the ceiling of its performance is still far away. Although the curves in Fig. 12 indicate that the default settings

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

10   IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING
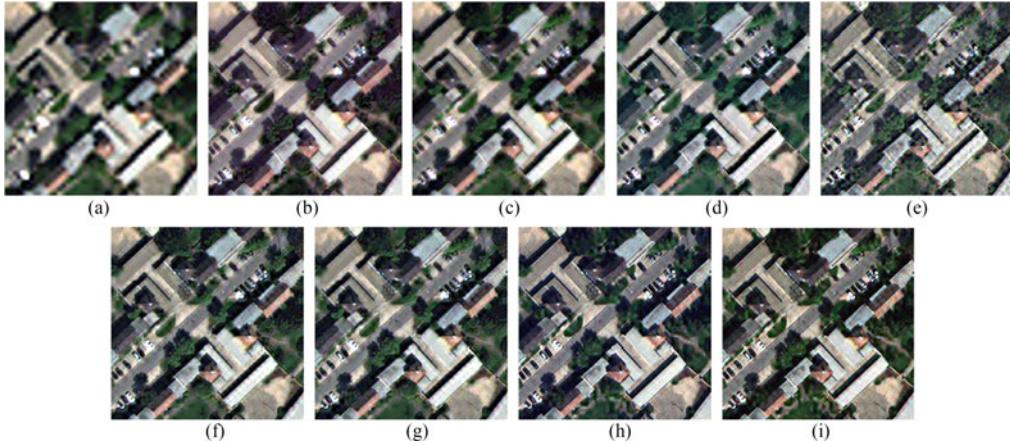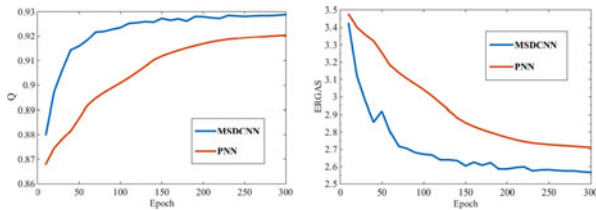

Fig. 11.   Results of the real-data experiment on an area of urban vegetation, which was extracted from a WorldView-2 image of San Francisco, United States, obtained in 2011. (a) Bicubic. (b) GS [4]. (c) PRACS [5]. (d) AWLP [42]. (e) TSSC [19]. (f) MTF-GLP [8]. (g) SFIM [10]. (h) PNN [33]. (i) MSDCNN.

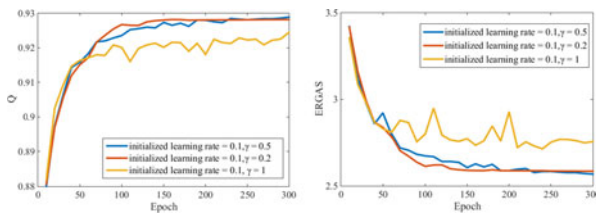Fig. 12.   Average Q and ERGAS of MSDCNN and PNN on the QuickBird dataset.

Fig. 13.   Average Q and ERGAS of MSDCNN with different values of γ.

work well, we tried another two settings for the learning rate γ to confirm our understanding of the learning process. The results of the comparison are shown in Fig. 13.

Fig. 13 helps us to confirm that the default setting of $\gamma = 0.5$ is a balanced decision between error decrease in the early training epochs and relatively smooth convergence in the later stages. Meanwhile, setting an appropriately low value for γ can lead to earlier convergence, but when γ is too small, the opportunity of breaking out of local minima may be lost.

2) Connection Architecture of the Multiscale Convolutional Layer Blocks: In the default architecture of MSDCNN, there is a flat convolutional layer between two multiscale blocks to reduce the spectral dimension from 60 to 30. To confirm its validity, two different architectures were compared, and their connections are illustrated in Fig. 14. In Block 2, two multiscale layers are contained in each block. In Block 3, a further skip connection is used, as in [27], and thus the spectral dimensionality is kept until the image is fed into the last layer.

By comparing the curves shown in Fig. 15, we can confirm that reducing the spectral dimensionality is necessary for the task, as the effect of using Block 3 without the reduction layer appears to be negative. From the comparison between Block 1
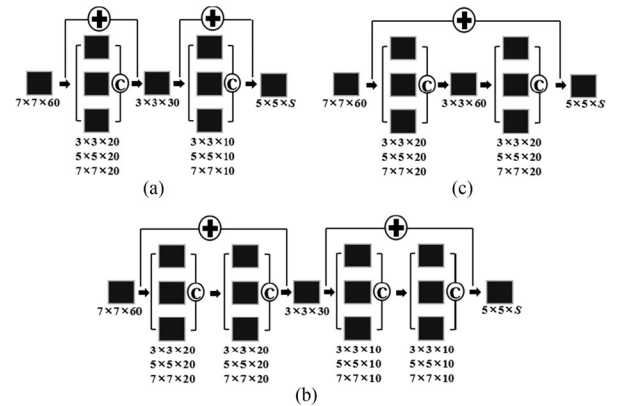

Fig. 14.   Block 1 is the architecture used in all the experiments undertaken in this study. (a) Block 1. (b) Block 2. (c) Block 3.
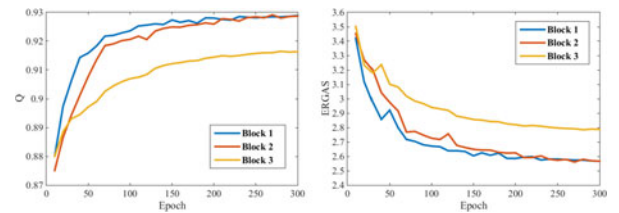

Fig. 15.   Average Q and ERGAS of MSDCNN with different values of γ on the QuickBird dataset.

and Block 2, we can observe that the deeper architecture needs more training epochs to reach a convergence region with a slightly higher accuracy, but such limited improvement is still far away from our expectation, and we assume that the network formed by Block 2 is not deep enough to fully develop the advantages of residual learning. Possible ways to reduce the training time cost will be studied in our future work.

V. CONCLUSION

In this paper, we have proposed a new CNN architecture for remote sensing imagery pan-sharpening. The main innovations in the model are the concepts of multiscale extraction, multidepth sharing, and merging of features from the spatial domain of the MS and PAN images. Compared with many of 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.

YUAN *et al.*: MULTISCALE AND MULTIDEPTH CONVOLUTIONAL NEURAL NETWORK FOR REMOTE SENSING IMAGERY PAN-SHARPENING 11

traditional and state-of-the-art pan-sharpening algorithms, the results of experiments undertaken on different datasets strongly indicate that the proposed MSDCNN is able to yield high-quality images with the best quantitative fidelity and appropriate sharpness.

In our future work, as the art of designing CNNs has not yet been fully explained from an analytical perspective instead of empirical ideas, there is still scope for the architecture of the proposed network to be optimized. Furthermore, our current feature learning strategies also require much study to transfer the obtained knowledge to some extended fields of remote sensing image fusion, quality improvement, and interpretation tasks, such as spatial-temporal unified fusion [50], hyperspectral image denoising [51], [52], aerial scene classification [53], [54], and target detection [55]. Furthermore, we also expect to develop advanced techniques of network compression and training data generalization, which helps to effectively process routine tasks on an application level.

References

[1] 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.

[2] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images .2. Channel ratio and chromaticity transformation techniques," *Remote Sens. Environ.*, vol. 22, pp. 343–365, Aug. 1987.

[3] P. S. Chavez and A. Y. Kwarteng, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, pp. 339–348, Mar. 1989.

[4] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," Google Patents US 6011875 A, 2000.

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

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

[7] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, pp. 143–156, Apr. 2007.

[8] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and pan imagery," *Photogramm. Eng. Remote Sens.*, vol. 72, pp. 591–596, May 2006.

[9] F. Palsson, J. R. Sveinsson, M. O. Ulfarsson, and J. A. Benediktsson, "MTF-based deblurring using a wiener filter for CS and MRA pansharpening methods," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 9, pp. 2255–2269, Jun. 2016.

[10] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, pp. 3461–3472, Dec. 2000.

[11] S. W. Zhong, Y. Zhang, Y. S. Chen, and D. Wu, "Combining component substitution and multiresolution analysis: A novel generalized BDSD pansharpening algorithm," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 10, no. 6, pp. 2867–2875, Jun. 2017.

[12] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, Jan. 2008.

[13] 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.

[14] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[15] L. Zhang, H. Shen, W. Gong, and H. Zhang, "Adjustable model-based fusion method for multispectral and panchromatic images," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 6, pp. 1693–1704, Dec. 2012.

[16] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[17] H. Shen, X. Meng, and L. Zhang, "An integrated framework for the spatio–temporal–spectral fusion of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7135–7148, Dec. 2016.

[18] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "A practical compressed sensing-based pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 4, pp. 629–633, Jul. 2012.

[19] C. Jiang, H. Zhang, H. Shen, and L. Zhang, "Two-step sparse coding for the pan-sharpening of remote sensing images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 7, no. 5, pp. 1792–1805, May 2014.

[20] 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.

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

[22] 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. 1, 2016.

[23] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3142—3155, Jul. 2017.

[24] P. Svoboda, M. Hradiš, L. Maršík, and P. Zemčík, "CNN for license plate motion deblurring," in *Proc. IEEE Int. Conf. Image Process.*, Phoenix, AZ, USA, 2016, pp. 3832–3836.

[25] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[26] N. Cai, Z. H. Su, Z. N. Lin, H. Wang, Z. J. Yang, and B. W. K. Ling, "Blind inpainting using the fully convolutional neural network," *Vis. Comput.*, vol. 33, pp. 249–261, Feb. 2017.

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

[28] J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 1637–1645.

[29] X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," *Adv. Neural Inf. Process. Syst.*, pp. 2802–2810, 2016.

[30] Y. Wang, L. Wang, H. Wang, and P. Li, "End-to-end image super-resolution via deep and shallow convolutional networks," arXiv:1607.07680, 2016.

[31] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

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

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

[34] A. Garzelli, "A review of image fusion algorithms based on the super-resolution paradigm," *Remote Sens.*, vol. 8, Oct. 2016, Art. no. 797.

[35] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Boston, MA, USA, 2015, pp. 1–9.

[36] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," Univ. Montreal, Montreal, QC, Canada, Tech. Rep. 1341, 2009.

[37] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," Presented at the Deep Learn. Workshop, Int. Conf. Machine Learn., 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[39] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, pp. 1–17, 1964.

[40] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[41] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB," *ACM Int. Conf. Multimedia*, pp. 689–692, 2015.

[42] X. Otazu, M. Gonzalez-Audicana, O. Fors, and J. Nunez, "Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 10, pp. 2376–2385, Oct. 2005.

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

12   IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING

[43] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. A. Licciardi *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, pp. 2565–2586, May 2015.

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

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

[46] L. Wald, *Data fusion: definitions and architectures: Fusion of images of different spatial resolutions*. Paris, France: Presses des MINES, 2002.

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

[48] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi-/hyper-spectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.

[49] L. Alparone, B. Alazzi, 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.

[50] P. Wu, H. Shen, L. Zhang, and F. M. Göttsche, "Integrated fusion of multi-scale polar-orbiting and geostationary satellite observations for the mapping of high spatial and temporal resolution land surface temperature," *Remote Sens. Environ.*, vol. 156, pp. 169–181, 2015.

[51] J. Li, Q. Yuan, H. Shen, and L. Zhang, "Noise removal from hyperspectral image with joint spectral-spatial distributed sparse representation," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5425–5439, Sep. 2016.

[52] Q. Yuan, L. Zhang, and H. Shen, "Hyperspectral image denoising employing a spectral-spatial adaptive total variation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 10, pp. 3660–3677, Oct. 2012.

[53] F. Hu, G. S. Xia, J. W. Hu, and L. P. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," *Remote Sens.*, vol. 7, pp. 14680–14707, Nov. 2015.

[54] G. S. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[55] F. Zhang, B. Du, L. Zhang, and M. Xu, "Weakly supervised learning based on coupled convolutional neural networks for aircraft detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 9, pp. 5553–5563, Sep. 2016.
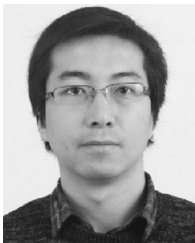
Qiangqiang Yuan (M'13) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is currently an Associate Professor. He published more than 50 research papers, including more than 30 peer-reviewed articles in international journals, such as the IEEE Transactions Image Processing and the IEEE Transactions on Geoscience and Remote Sensing. His current research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was the recipient of the Top-Ten Academic Star of Wuhan University, in 2011 the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council, in 2014. He has frequently served as a Referee for more than 20 international journals for remote sensing and image processing.

Yancong Wei (SM'16) received the B.S. degree in geodesy and geomatics engineering from Wuhan University, Wuhan, China, in 2015. Since 2015, he has been working toward the M.S. degree at the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University.

His research interests include degraded information reconstruction for remote sensed images, data fusion, and computer vision.

Mr. Wei was the recipient of Academic Scholarship for Undergraduate Students, in 2014, and Academic Scholarship for Graduate Students, in 2017, awarded by Wuhan University.
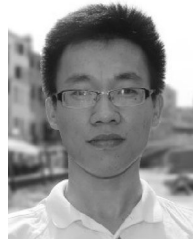
Xiangchao Meng received the B.S. degree in geographic information system from Shandong University of Science and Technology, Qingdao, China, in 2012, the Ph.D. degree in cartography and geography information system from Wuhan University, Wuhan, China, in 2017.

He is currently a Lecturer with the Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo, China. His research interests include remote sensing image fusion and quality evaluation.
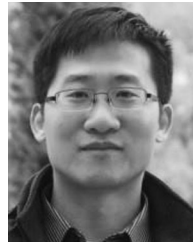
Huanfeng Shen (M'10–SM'13) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2002 and 2007, respectively.

In 2007, he joined the School of Resource and Environmental Sciences, Wuhan University, where he is currently a Luojia Distinguished Professor. He has authored more than 100 research papers. His research interests include image quality improvement, remote sensing mapping and application, data fusion and assimilation, and regional and global environmental change.

Dr. Shen is currently a member of the editorial board of the *Journal of Applied Remote Sensing*. He has been supported by several talent programs, such as the Youth Talent Support Program of China (2015), China National Science Fund for Excellent Young Scholars (2014), and the New Century Excellent Talents by the Ministry of Education of China (2011).

Liangpei Zhang (M'06–SM'08) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He is currently the Head of the remote sensing division, State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He is also a "Chang-Jiang Scholar" Chair Professor appointed by the ministry of education of China. He is currently a Principal Scientist for the China State Key Basic Research Project (2011–2016) appointed by the Ministry of National Science and Technology of China to lead the remote sensing program in China. He has authored or coauthored more than 500 research papers and five books. He is the holder of 15 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is the Founding Chair of IEEE Geoscience and Remote Sensing Society (GRSS) Wuhan Chapter. He was the recipient of the best reviewer awards from IEEE GRSS for his service to IEEE Journal of Selected Topics in Earth Observations and Applied Remote Sensing (JSTARS), in 2012, and IEEE Geoscience and Remote Sensing Letters, in 2014, and also the 2010 best paper Boeing award and the 2013 best paper ERDAS award from the American Society of Photogrammetry and Remote Sensing. He was the General Chair for the 4th IEEE GRSS Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS) and the Guest Editor for JSTARS. His research teams was the recipient of the top three prizes of the IEEE GRSS 2014 Data Fusion Contest, and his students have been selected as the winners or finalists of the IEEE International Geoscience and Remote Sensing Symposium student paper contest in recent years. He is a Fellow of the Institution of Engineering and Technology (IET), Executive Member (board of governor) of the China National Committee of International Geosphere–Biosphere Programme, Executive Member of the China Society of Image and Graphics, etc. He regularly serves as the Cochair of the series SPIE Conferences on Multispectral Image Processing and Pattern Recognition, Conference on Asia Remote Sensing, and many other conferences. He edits several conference proceedings, issues, and geoinformatics symposiums. He also serves as an Associate Editor for the *International Journal of Ambient Computing and Intelligence, International Journal of Image and Graphics, International Journal of Digital Multimedia Broadcasting, Journal of Geo-spatial Information Science, and Journal of Remote Sensing,* and the Guest Editor for *Journal of Applied Remote Sensing and Journal of Sensors.* He is currently serving as an Associate Editor of the IEEE Transactions on Geoscience and Remote Sensing.