Pansharpening via Detail Injection Based Convolutional Neural Networks

Lin He, *Member, IEEE*, Yizhou Rao, Jun Li, *Senior Member, IEEE*, Jocelyn Chanussot, *Fellow, IEEE*, Antonio Plaza, *Fellow, IEEE*, Jiawei Zhu, and Bo Li

***Abstract*—Pansharpening aims to fuse a multispectral (MS) image with an associated panchromatic (PAN) image, producing a composite image with the spectral resolution of the former and the spatial resolution of the latter. Traditional pansharpening methods can be ascribed to a unified detail injection context, which views the injected MS details as the integration of PAN details and bandwise injection gains. In this paper, we design a new detail injection based convolutional neural network (DiCNN) framework for pansharpening with the MS details being directly formulated in end-to-end manners, where the first detail injection based CNN (DiCNN1) mines MS details through the PAN image and the MS image, and the second one (DiCNN2) utilizes only the PAN image. The main advantage of the proposed DiCNNs is that they provide explicit physical interpretations and can achieve fast convergence while achieving high pansharpening quality. Furthermore, the effectiveness of the proposed approaches is also analyzed from a relatively theoretical point of view. Our methods are evaluated via experiments on real MS image datasets, achieving excellent performance when compared to other state-of-the-art methods.**

***Index Terms*—Convolutional neural networks (CNNs), detail injection, pansharpening.**

Manuscript received December 9, 2018; revised January 27, 2019; accepted January 31, 2019. This work was supported in part by the National Natural Science Foundation of China under Grants 61571195, 61771496, 61633010, and 61836003, in part by the Guangdong Provincial Natural Science Foundation under Grants 2016A030313254, 2016A030313516, and 2017A030313382, in part by the National Key Research and Development Program of China under Grant 2017YFB0502900, and in part by China Scholarship Council under Grant 201706155080. *(Corresponding author: Jun Li.)*

L. He, Y. Rao, and J. Zhu are with the School of Automation Science and Engineering, South China University of Technology, Guangzhou 510640, China (e-mail: helin@scut.edu.cn; raoyizhou@139.com; zjw7342@163.com).

J. Li is with the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: lijun48@mail.sysu.edu.cn).

J. Chanussot is with the University of Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

A. Plaza is with the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, Escuela Politécnica, University of Extremadura, Cáceres E-10071, Spain (e-mail: aplaza@unex.es).

B. Li is with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, and the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: boli@buaa.edu.cn).

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.2019.2898574

I. Introduction

Due to the physical characteristics of multispectral (MS) image sensors, they generally acquire MS images with limited spatial resolution. However, high spatial resolution MS images are required in many applications, such as classification, target detection, scene interpretation, and spectral unmixing [1], [2]. Therefore, pansharpening has been an active area of research, drawing significant attention in the area of remotely sensed image processing. The pansharpening task aims at fusing a low spatial resolution MS image and a registered wide-band panchromatic (PAN) image, utilizing the detail information contained in the PAN image to sharpen the MS image, hence yielding a high spatial resolution MS image [1]. The task can be seen as a special reconstruction based on different types of data with different characteristics. For simplicity, low spatial resolution MS images are called LRMS images, and high spatial resolution MS images are called HRMS images, hereinafter. A HRMS image pansharpened from the LRMS image is called pansharpened HRMS image hereinafter. Ideally, as a full resolution image, the pansharpened HRMS image should have the same spectral resolution as the original LRMS image and the same spatial resolution as the corresponding PAN image.

Over the past decades, a wide variety of pansharpening methods have been proposed in the literature [1]–[4]. Among such existing methods, component substitution (CS) and multi-resolution analysis (MRA) are two widely representative categories [1]–[3]. CS approaches usually replace certain components of the MS image with those from the PAN image in a given domain, which include principal component analysis based pansharpening [5]–[7], Brovey transform based pansharpening [8], [9], and Gram–Schimidt (GS) transform based pansharpening [10], [11], among others. In contrast, MRA methods exploit the spatial information via a multiresolution decomposition of the images, which generally involves detail extraction and detail integration in multiple scales. Examples are pansharpenings based on decimated wavelet transform [12], undecimated wavelet transform [13], *a tróus* wavelet transform (ATWT) [14]–[16], and Laplacian pyramid [17]–[19]. The aforementioned methods differ mainly in how spatial details are extracted from the PAN image and how they are injected into the preinterpolated LRMS image. One major challenge for CS/MRA approaches is to preserve spatial details resolved from the PAN image as much as possible, while avoiding spectral distortion. This refers to the spectral deviation from an ideal spectrum, especially when PAN and MS images are acquired in spectral ranges that overlap only partially [1], [20]. Unfortunately, existing CS/MRA methods are often prone to significant spectral distortion [3], even under some improvement of fusion strategies such as histogram matching [21], weighted detail injection [16], or some hybrid intermediate processes [22]. This is probably due to the fact that the details are not very effectively learned and injected, although CS/MRA approaches indeed aim to utilize the detail information.

Recently, convolutional neural networks (CNNs) start prevailing in image enhancement tasks such as super-resolution [23], [24] and pansharpening [20], [25]. Super-resolution is, to some degree, a pansharpening-related task, as both super-resolution and pansharpening aim to enhance image resolution. There are, however, some differences among them since the former is usually a single input single output (SISO) process while the latter is a multiple input single output (MISO) case. Dong *et al.* proposed a super-resolution CNN (SRCNN), which is a three-layer CNN, to learn the mapping from the input low-resolution image to the output high-resolution image [23]. Kim *et al.* designed a deep CNN structure for super-resolution, where the residual component is learned [26]. Whether or not details are injected from the PAN image to its associated LRMS image represents the major difference between pansharpening and super-resolution tasks. Considering this, Masi *et al.* presented a pansharpening CNN (PNN) following the basic thread of SRCNN [20], where the pre-interpolated LRMS image is stacked with the PAN image at the input layer, and then a CNN process is used to learn the relationship between the input and the pansharpened HRMS image. Although PNN exhibits good performance on real remotely sensed data, difficulties arise from the long-time training iterations and the problem that it misses the domain specific pansharpening structure and roughly treats pansharpening as a black-box learning procedure. Afterwards, Wei *et al.* designed a CNN method for pansharpening [25]. The method comprises the process of residual learning and the subsequent dimension reduction, which is faced with the problems that the learned residual has no explicit physical interpretation for pansharpening and there is an additional computation load related to dimension reduction. They also introduce strategies such as multiscale kernels into the CNN-based pansharpening [27].

In this paper, a general detail injection formulation, namely, detail injection based CNNs for pansharpening (DiPAN), is summarized, which is able to accommodate CS/MRA pansharpening methods. The proposed DiPAN can be used as a domain-specific structure to guide the design of new pansharpening methods. In the context of our DiPAN framework, two detail injection based CNNs (DiCNNs) for MS detail learning are introduced, where the main contributions of this paper can be summarized as follows.

1) The first method, called DiCNN1, adopts a framework in which the pathway of stacked convolutional layers only learns the MS details from the combination of the pre-interpolated LRMS image and the PAN image in an end-to-end manner, resulting in good initialization. DiCNN1, following the basic idea in our previous work [28], has clear interpretability in the detail injection context, and can greatly reduce the uncertainty of learning, thus achieving high computational efficiency and pansharpening quality. A detailed description of the method, followed by a discussion and extensive experimental results, are provided in this paper. Furthermore, we present a theoretical analysis and proof of the effectiveness of DiCNN1. To the best of our knowledge, the effectiveness of a parsharpening CNN has not been previously explored from such a theoretical point of view.
2) The second method, called DiCNN2, is capable of transfer learning when there are bad bands in test MS images. DiCNN2 works under the assumption that ideal MS detail is only relevant to the PAN image, and directly uses the PAN image as the input of the convolutional layer pathway, which makes it able to perform transfer learning in addition to the regular pansharpening task. Since its input is a one-dimensional (1-D) PAN image only (with a small amount of CNN free parameters), DiCNN2 yields very fast computation.

The remainder of the paper is organized as follows. Section II introduces the detail injection framework. Section III summarizes major existing CNN-based super-resolution and pansharpening methods. Section IV introduces our detail injection based CNN pansharpening methods and presents the corresponding complexity analysis. Section V evaluates the proposed methods via experiments with real MS datasets. Section VI concludes the paper with some remarks and hints at plausible future research lines.

II. Detail Injection Framework

Let $\mathbf{P} \in \mathbb{R}^{H\times W}$ denote an observed PAN image with size $H \times W$; let $\widetilde{\mathbf{M}} \in \mathbb{R}^{H\times W\times N_b}$ be a pre-interpolated LRMS, which has been interpolated spatially to the scale of the PAN image (with N_b being the number of bands); and let $\widehat{\mathbf{M}}$ be the pansharpened HRMS image.

Traditionally, CS/MRA methods are viewed as two major groups of pansharpening methods [1]. CS category can be generally formulated as

$$\widehat{\mathbf{M}}_b = \widetilde{\mathbf{M}}_b + g_b \cdot (\mathbf{P} - \mathbf{I}_c), \quad b = 1, \ldots, N_b \tag{1}$$

where $\widehat{\mathbf{M}}_b$ and $\widetilde{\mathbf{M}}_b$ are the bth bands of $\widehat{\mathbf{M}}$ and $\widetilde{\mathbf{M}}$, respectively, g_b represents the injection gain associated with $\widetilde{\mathbf{M}}_b$, N_b is the number of MS bands, and $\mathbf{I}_c$ is the intensity component of the MS image, which is often a weighted sum $\mathbf{I}_c = \sum_{b=1}^{N_b} \omega_b \widetilde{\mathbf{M}}_b$. To show the substitution process in CS methods, (1) can be reformulated as

$$\begin{aligned}\widehat{\mathbf{M}}_b &= \widetilde{\mathbf{M}}_b - \mathbf{I}_c + g_b \cdot (\mathbf{P} - \mathbf{I}_c) + \mathbf{I}_c \\ &= (\widetilde{\mathbf{M}}_b - \mathbf{I}_c) + g_b \cdot \left(\mathbf{P} - \frac{g_b - 1}{g_b}\mathbf{I}_c\right)\end{aligned} \tag{2}$$

which suggests that, in a CS method, the component $\mathbf{I}_c$ is substituted with the component $g_b \cdot (\mathbf{P} - \frac{g_b-1}{g_b}\mathbf{I}_c)$. On the other hand, the general formulation of MRA methods is of the form [1]

$$\widehat{\mathbf{M}}_b = \widetilde{\mathbf{M}}_b + g_b \cdot (\mathbf{P} - \mathbf{P}_c), \quad b = 1, \ldots, N_b \tag{3}$$

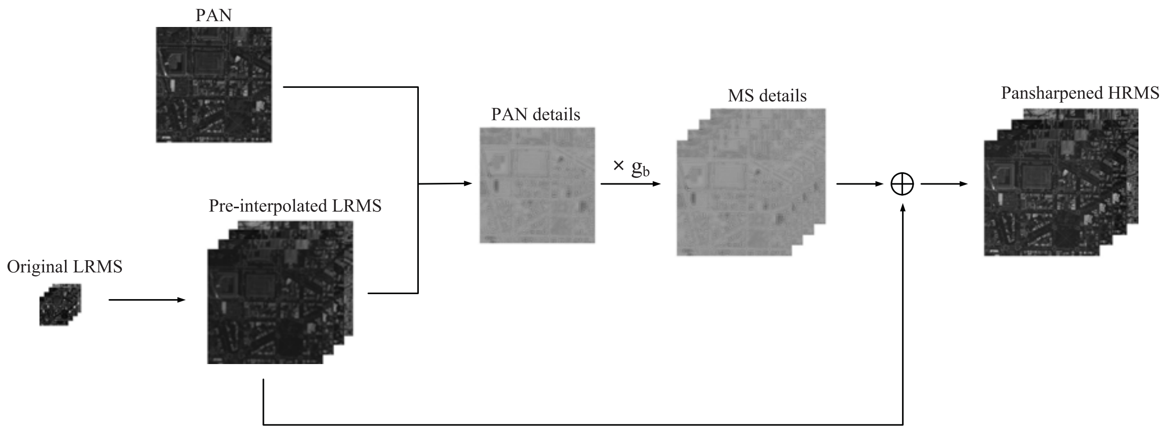


Fig. 1. Schematic diagram of the DiPAN framework.

where $\mathbf{P}_c$ denotes the low-frequency component of the PAN image, which is usually obtained in a MRA way. According to the representations in (1) and (3), both CS and MRA methods are normally based on two sequential phases: First, the extraction of MS details from the PAN image, which usually comprises intermediate processes, such as yielding PAN details and obtaining band injection gains. Second, the injection of the MS details into the LRMS image to produce the HRMS image. Therefore, such two categories of pansharpening methods can be represented in a unified detail injection framework, namely DiPAN, as follows:

$$\begin{aligned}\widehat{\mathbf{M}}_b &= \widetilde{\mathbf{M}}_b + g_b \cdot \mathbf{d} \\ &= \widetilde{\mathbf{M}}_b + \mathbf{D}_b \end{aligned} \tag{4}$$

where $\mathbf{d}$ represents the PAN details, which are usually calculated by involving both the PAN image and the MS image with a certain criterion, $\mathbf{D}_b = g_b \cdot \mathbf{d}$ denotes the MS details, which should complement the pre-interpolated LRMS image $\widetilde{\mathbf{M}}$, while g_b stands for the associated injection gain responsible for transferring the PAN details to the MS details. A schematic diagram of DiPAN is given in Fig. 1, where it is indicated that the full-resolution pansharpened HRMS image $\widehat{\mathbf{M}}$ can be decomposed into the MS details and the LRMS approximation.

As the formulation in (4) and the schematic diagram in Fig. 1 indicate, DiPAN has clear physical interpretability for the pansharpening process, which can be used as a pansharpening domain-specific structure to guide the design of new pansharpening methods.

III. Super-Resolution and Pansharpening Using CNN Strategy

Recently, CNNs were successfully applied in image super-resolution and pansharpening. CNNs are usually treated as the descendants of traditional artificial neural networks [29]–[31], in which assumptions such as a limited receptive field (processing input only in a neuron's local neighborhood) and the spatial invariant weight (so-called weight sharing) are normally jointly employed.

The response of a convolutional layer in a CNN can be given by

$$\mathbf{Y}_l = \varphi(\mathbf{W}_l * \mathbf{X}_l + \mathbf{B}_l) \tag{5}$$

where $*$ denotes the convolution operation, $\mathbf{X}_l$ and $\mathbf{Y}_l$ are the input and output of the lth layer, respectively, $\mathbf{W}_l$ and $\mathbf{B}_l$ are the weight and bias metrics, respectively, and $\varphi(\cdot)$ represents the activation function. Due to its ability to mitigate gradient vanishing and its computational simplicity, the rectified linear unit (ReLU) [32] is commonly used in CNNs, whose input–output relation is $\mathbf{Y}_l = \max(0, \mathbf{X}_l)$ [23], [33]–[35].

Both image super-resolution and pansharpening intend to recover high-resolution images from the observed low-resolution data, with the major disparity being that one is a SISO process and the other one is a MISO one. In image super-resolution, usually the low spatial resolution image (as a single input) is processed to output a high spatial resolution image, while pansharpening utilizes the MS image with low spatial resolution and the PAN image with low spectral resolution as two separate data sources to recover the full resolution HRMS image. The two kinds of image resolution enhancements mentioned above are used as mathematical tools to minimize the loss function of expected square error as

$$\ell(\boldsymbol{\theta}) = E\|\widehat{\mathbf{H}}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|_F^2 \tag{6}$$

where $\widehat{\mathbf{H}}$ is the predicted high-resolution image following a parametric structure, $\mathbf{Y}$ is the ideal high-resolution image, $\boldsymbol{\theta}$ denotes the parameters used to infer the predicted image, and $\mathbf{X}$ is the low-resolution input, which means a low spatial resolution image for image super-resolution that represents both the low spectral resolution PAN image and the associated LRMS image for pansharpening.

Dong *et al.* designed a three-layer CNN for image super-resolution able to directly learn the mapping between the low-resolution image and the high-resolution image, which is called super-resolution CNN (SRCNN) [23]. Therein patch extraction and representation are used to improve computational efficiency and feature locality in the training phase. The objective is to minimize the following patchwise mean square error:

$$\ell(\boldsymbol{\theta}) = E\|\widehat{\mathbf{H}}(\mathbf{X};\boldsymbol{\theta}) - \mathbf{Y}\|_F^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\|\widehat{\mathbf{H}}^{(i)}(\mathbf{X}^{(i)};\boldsymbol{\theta}) - \mathbf{Y}^{(i)}\|_F^2 \tag{7}$$

where i is the index of patches, N_p denotes the number of total patches, $\boldsymbol{\theta}$ represents the free CNN parameters to be optimized under the CNN context, $\mathbf{X}^{(i)}$ refers to the ith patch in the low-resolution image, and $\widehat{\mathbf{H}}^{(i)}$ stands for the ith patch in the predicted high-resolution image. As a counterpart for pansharpening purposes, Masi *et al.* introduced a PNN [20], which stacks the pre-interpolated LRMS image and the PAN image together and then uses a CNN to mine the mapping between this concatenation and a real HRMS image.

The loss function to be minimized is

$$\ell(\boldsymbol{\theta}) = E\|\widehat{\mathbf{M}}(\mathbf{G};\boldsymbol{\theta}) - \mathbf{Y}\|_F^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\|\widehat{\mathbf{M}}^{(i)}(\mathbf{G}^{(i)};\boldsymbol{\theta}) - \mathbf{Y}^{(i)}\|_F^2 \tag{8}$$

where $\mathbf{G} = (\widetilde{\mathbf{M}}, \mathbf{P})$ in the size $H \times W \times (N_b + 1)$ denotes the concatenation of the pre-interpolated LRMS image $\widetilde{\mathbf{M}}$ and the PAN image $\mathbf{P}$ along the band dimension. Here, the target $\mathbf{Y}$ stands for the ideal HRMS for the pansharpening case. Considering that MS images are in 3-D data arrangement, $\widehat{\mathbf{M}}$ and $\mathbf{Y}$ are originally three-way or third-order tensors [36]. To accommodate a matrix representation, $\widehat{\mathbf{M}}$ and $\mathbf{Y}$ in (8) are unfolded as matrices, for example, along the first mode and being denoted as $\widehat{\mathbf{M}}_{(1)}$ and $\mathbf{Y}_{(1)}$ [36]. But, for simplicity, $\widehat{\mathbf{M}}$ and $\mathbf{Y}$ in (8) represent their unfolding matrices $\widehat{\mathbf{M}}_{(1)}$ and $\mathbf{Y}_{(1)}$, respectively. If not stated otherwise, the reminder of the paper follows the same expression routine when involving three-way tensor representations.

The deep residual network (ResNet) has reached excellent performance in image classification [37]. Its success largely stems from attaching an identity skip connection to fit a residual mapping. Kim *et al.* extended ResNet and proposed a deep network for super-resolution purposes, which intends to learn the residual supplementary to the input low-resolution image instead of the predicted high-resolution image itself [26]. The loss function is defined as follows:

$$\ell(\boldsymbol{\theta}) = E\|\widehat{\mathbf{R}}(\mathbf{X};\boldsymbol{\theta}) + \mathbf{X} - \mathbf{Y}\|_F^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\|\widehat{\mathbf{R}}^{(i)}(\mathbf{X}^{(i)};\boldsymbol{\theta}) + \mathbf{X}^{(i)} - \mathbf{Y}^{(i)}\|_F^2 \tag{9}$$

where $\mathbf{R}$ represents the residual that needs to be learnt. Later, Wei *et al.* used a similar strategy for pansharpening, termed deep residual pansharpening neural network (DRPNN) [25]. In the DRPNN, the concatenation of the pre-interpolated LRMS image and the PAN image pass through both stacked layers and a shortcut connection to yield the residual and, then, an additional convolutional layer is included for dimensionality reduction. The connected objective is to minimize the following loss:

$$\ell(\boldsymbol{\theta}) = \|\omega(\widehat{\mathbf{R}}(\mathbf{G};\boldsymbol{\theta}) + \mathbf{G}) - \mathbf{Y}\|_F^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}\|\omega(\widehat{\mathbf{R}}^{(i)}(\mathbf{G}^{(i)};\boldsymbol{\theta}) + \mathbf{G}^{(i)}) - \mathbf{Y}^{(i)}\|_F^2 \tag{10}$$

where $\omega(\cdot)$ denotes a convolution operation for dimensional matching.

In comparison with the CS/MRA approaches, CNNs provide a new possibility to perform learning for pansharpening, where the details are driven from the context. However, in comparison with DiPAN, the main limitation of the aforementioned CNN-based pansharpening approaches is the lack of physical interpretability, and the fact that they do not use an appropriate domain-specific structure. The weaknesses are, specifically, as follows.

1) PNN treats pansharpening merely as a black-box learning procedure, without considering the domain-specific structure useful to pansharpening, which results in a heavy training process and limited learning ability.
2) DRPNN involves the structure of residual and the subsequent dimension reduction, which faces the problem that the processed residual has no explicit physical interpretation in a pansharpening context, and there is additional computational burden for the dimension reduction step.

IV. Proposed Methods

Based on the DiPAN framework in Section II, we develop DiCNNs for pansharpening. The advantages of the proposed DiCNNs can be summarized as follows.

1) We take into consideration the detail structure used in traditional CS/MRA-based pansharpening and then directly learn MS details, without separating the PAN details and the connected gains. This allows us to circumvent the intermediate process needed to learn such two pieces of information individually, thus reducing the model uncertainty.
2) Compared to existing CNN-based pansharpening methods, our newly proposed methods have clear and meaningful interpretation in the context of detail injection and can also achieve excellent learning performance.

A. First Detail Injection Based CNN (DiCNN1)

Following DiPAN, our pansharpening method focuses on reconstructing the MS details in a CNN manner. To achieve this goal, we build a feedforward neural network, where a shortcut connection skips three stacked convolutional layers and the output of the shortcut is added to the output of stacked layers to yield the predicted HRMS [as shown in Fig. 2(a)]. This network employs the concatenation of the pre-interpolated LRMS and the PAN images as the input. However, only the pre-interpolated LRMS is propagated through the shortcut connection. In this way, the stacked layers utilize the interaction of the pre-interpolated LRMS and PAN images to yield only the MS details that can further supplement the LRMS

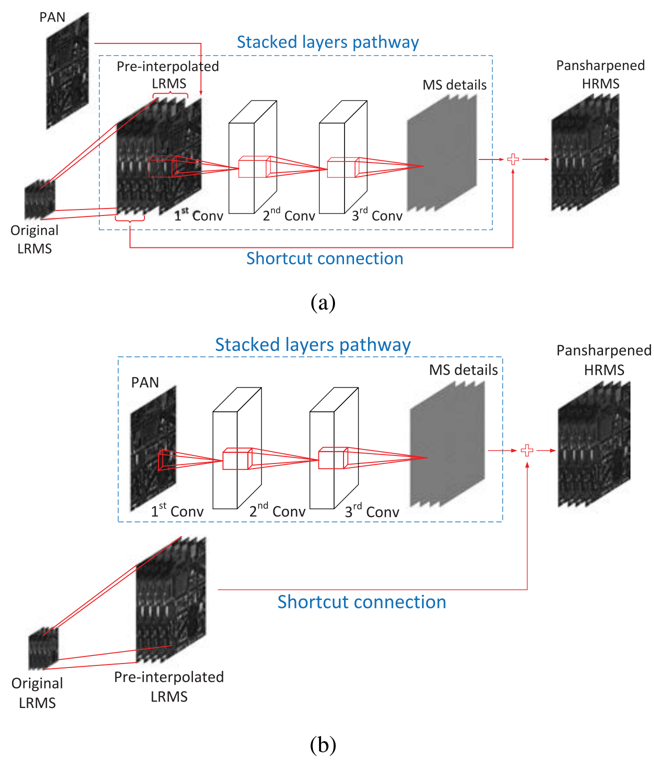


Fig. 2. Architectures of (a) DiCNN1 and (b) DiCNN2.

image in order to produce the pansharpened HRMS image. Specifically, our objective is to minimize the following loss function:

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \|\widehat{\mathbf{D}}(\mathbf{G};\boldsymbol{\theta}) + \widetilde{\mathbf{M}} - \mathbf{Y}\|_F^2 \\ &= \frac{1}{N_p}\sum_{i=1}^{N_p} \|\widehat{\mathbf{D}}^{(i)}(\mathbf{G}^{(i)};\boldsymbol{\theta}) + \widetilde{\mathbf{M}}^{(i)} - \mathbf{Y}^{(i)}\|_F^2 \qquad (11)\end{aligned}$$

where $\widehat{\mathbf{D}}$ represents the MS details reconstructed with the input $\mathbf{G}$, the concatenation of the LRMS image and the PAN image, and the parameter $\boldsymbol{\theta}$.

Practically, pansharpening is an ill-posed problem, which means that many solutions exist for a given low-resolution input. This is mathematically connected to an underdetermined inverse problem, of which the solution is not unique. In theory, such a problem can be relieved by constraining the solution space with appropriate prior information, which influences the overall performance of pansharpening. Fig. 3 depicts the basic structure of several CNN-based methods, with Fig. 3(a) and (b) representing the PNN and DRPNN (mentioned previously) and Fig. 3(c) representing our DiCNN1. As we can observe, the PNN directly learns the mapping between its input (the pre-interpolated LRMS image plus the PAN image) and the reconstructed HRMS image, without involving any prior knowledge on structure, regarding pansharpening just as a black-box learning problem. In the DRPNN, a residual structure is introduced into pansharpening [as shown in Fig. 3(b)], motivated by the residual learning process for image super-resolution in [26]. However, this residual structure brings some inherent weaknesses when used for pansharpening. First, DRPNN uses the concatenation of the pre-interpolated LRMS image and the PAN image as its input. This input goes through the stacked layers and the shortcut connections simultaneously, which forces the output of stacked layers pathway to be of the same dimensionality as the input of the input concatenation, i.e., one dimension more than that of the pansharpened HRMS image, thus yielding a residual learning result that has no explicit physical interpretation in a pansharpening context. Second, this dimensionality mismatch has to rely on an extra convolutional layer, which apparently aggravates the computational burden.

Different from PNN and DRPNN, our DiCNN1 takes into consideration the structure of the detail injection framework. It uses the concatenation of the pre-interpolated LRMS image and the PAN image as the input of the stacked layers, whereas the shortcut connection inputs only the pre-interpolated LRMS image. This strategy makes the output of stacked layers pathway be the MS details that can directly supplement the pre-interpolated LRMS image to produce the HRMS image, which guarantees that this CNN is able to directly learn the MS details. This implies that DiCNN1 does introduce a domain-specific structure with meaningful interpretation, meanwhile excluding the additional computational burden. On the other hand, compared to the detail injection based CS and MRA methods, DiCNN1 learns only the MS details *per se*, avoiding to separately process the PAN details and the associated gains and, hence, reducing the model uncertainty.

B. *Second Detail Injection Based CNN (DiCNN2)*

When a PNN model has been trained, the test MS images may be changed; for example, bad bands may be present in the data. In this situation, can a PNN model be transferred to pansharpen those different kinds of test images?

As mentioned in previous sections, pansharpening utilizes the details mainly existing in the PAN image to supplement the LRMS image, so as to obtain the HRMS image. These details can be viewed as the result from a filtering process, where certain low-frequency components are filtered out [38], which is a common rule for pansharpening on various sorts of images. Under this rule, it therefore makes sense that, for a given CNN, different sets of parameters suitable for pansharpening different kinds of images have certain inherent connections. As a result, it is possible to use a pre-trained CNN model on a certain kind of images for pansharpening other kinds of images. This is actually a transfer learning process [39]. By closely inspecting Fig. 2(a), we can see that both the PAN image and the LRMS image are fed into the convolutional layers pathway, which indicates that the LRMS image will significantly affect the extraction of details when the type of the MS image varies and, thus, reduce the robustness of the model learning in the stack layers pathway. To address this issue, we have developed another PNN, called DiCNN2 [as shown in Fig. 2(b)]. In DiCNN2, only the PAN image is connected to the convolutional layers pathway, which removes the influence of the LRMS image on detail extraction. Though this may also reduce the specificity of details for certain kinds of MS images, the shortcut connection still inputs 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.

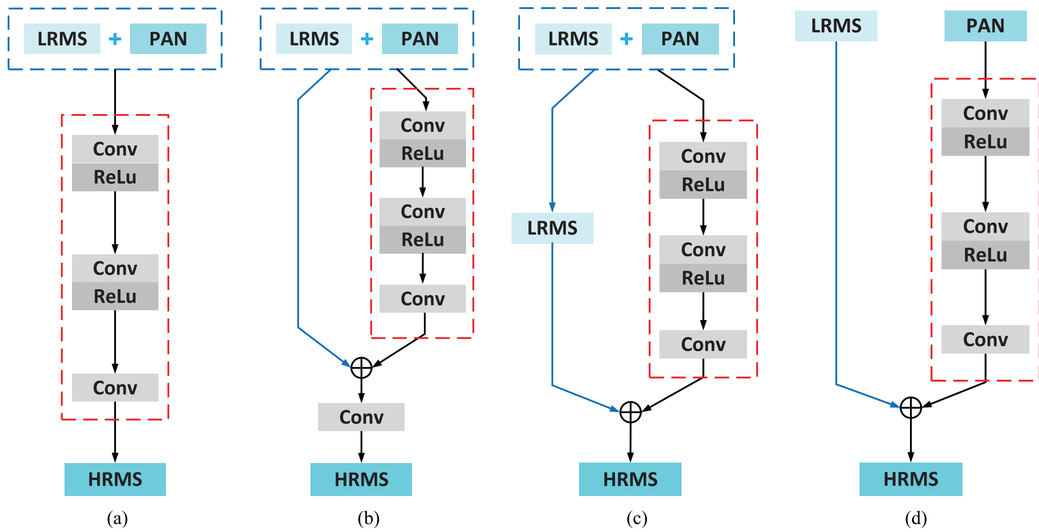


Fig. 3. Structural comparison of (a) PNN, (b) DRPNN, (c) DiCNN1, and (d) DiCNN2, where the red dashed-line box marks the convolutional layers pathway and $\bigoplus$ represents a pixelwise addition.

pre-interpolated LRMS image to force the convolutional layers pathway to learn only the information about the MS details. The objective function to be minimized for DiCNN2 is

$$\ell(\boldsymbol{\theta}) = \|\widehat{\mathbf{D}}(\mathbf{P};\boldsymbol{\theta}) + \widetilde{\mathbf{M}} - \mathbf{Y}\|_F^2$$

$$= \frac{1}{N_p}\sum_{i=1}^{N_p} \|\widehat{\mathbf{D}}^{(i)}(\mathbf{P}^{(i)};\boldsymbol{\theta}) + \widetilde{\mathbf{M}} - \mathbf{Y}^{(i)}\|_F^2. \qquad (12)$$

In real applications, once a CNN is trained, the network parameters in the convolutional layers pathway are fixed, except for those on the last layer. When a new kind of images are input, only this layer needs to be fine-tuned.

It is noteworthy that DiCNN2 is also a kind of detail injection based CNN. In addition to performing pre-training transfer, DiCNN2 can be seen as an alternative to DiCNN1 for usual pansharpening tasks, where the data for training and prediction come from the same sensors. Fig. 3(d) depicts the simplified structure of such a PNN, which suggests that DiCNN2 can provide similar benefits as DiCNN1, such as meaningful detail injection interpretation, high computational efficiency, and model simplification. Especially, DiCNN2 uses the PAN image as the input of the stacked convolutional layers, in contrast with the concatenation of the PAN image and multi-band LRMS image, thus leading to even higher computational efficiency than DiCNN1.

In summary, we developed DiCNNs, both of which exhibit the capacity to perform the regular pansharpening task. But they also differ in the following three main aspects.

1) DiCNN1 and DiCNN2 have different network structures. As it can be observed in Fig. 2(a) and (b), and Fig. 3(c) and (d), DiCNN1 inputs both the PAN image and the LRMS to the convolutional pathway, whereas DiCNN2 inputs only the PAN image to the convolutional pathway.
2) DiCNN1 and DiCNN2 are connected to different loss functions. This is because the different structures of the two CNNs lead to such different functions. Specifically, the loss function of DiCNN1 is given by (11), while that of DiCNN2 is given by (12). The convolutional kernels that need to be resolved in DiCNN1 are coupled with the concatenation of both the MS image and the PAN image, but in DiCNN2 those kernels are convolved only with the PAN image.
3) DiCNN1 and DiCNN2 can serve different purposes. Both DiCNN1 and DiCNN2 are able to perform the regular pansharpening task. However, DiCNN2 exhibits an additional ability to perform transfer learning. In DiCNN2, the LRMS image is separated out from the input of the convolutional pathway, which means that the bottom layer of the CNN's convolutional pathway is not tightly relevant to it. Thus, for a test LRMS image with bad bands, we only need to fine-tune the top convolutional layer, which is a kind of transfer learning.

V. Experimental Results

This section evaluates the performance of our pansharpening methods, where three real remotely sensed image datasets are considered. These datasets were acquired with WorldView-2, IKONOS, and Quickbird sensors. During the evaluation, we conduct reduced-resolution and full-resolution experiments, as well as transfer learning experiments.

In the case of reduced-resolution assessments, we set experiments using Wald's protocol [40]. The MS image and the PAN image were degraded to a lower resolution by using a Gaussian filter with a factor of 4 [41], and then the degraded MS image was pre-interpolated to the same spatial size as the degraded PAN image using a polynomial kernel (EXP) [4]. The criteria used for the assessment include x-band extension of universal image quality index (Qx) [42], spatial correlation coefficient (SCC) [43], spectral angle mapper (SAM) [44], and *Erreur Relative Globale Adimensionnelle de Synthèse* (ERGAS) [45]. These indexes are widely used to measure the qualities of pansharpened images, with the original MS image as the ground-truth.

TABLE I
COMPARISON AMONG CNN-BASED METHODS

	PNN	DRPNN	DiCNN1	DiCNN2
Detail Learning	No	No	Yes	Yes
Residual Learning	No	Yes	No	No
Transfer Learning	No	No	No	Yes

For fair comparison, we apply consistent parameter setting to different CNN-based pansharpening methods. Specifically, the number of convolutional layers in the convolutional pathway for all PNNs in comparison are set to be 3. Thus, we can compare the results under the basic network structure while avoiding the influence of the deepness of the hidden layers. In addition, each of them utilizes 64 filters with spatial size 3×3, except for the last layer with N_b filters. Furthermore, all the training patches and validation patches are with the spatial size 41×41, totally being 25 600 patches, wherein 64 patches are randomly selected from training data as a mini-batch for SGD. The number of training iterations is set to 3.0×10^5 in all cases. The learning rate is initially set to 0.0001 and it adaptively updates based on Adam[46]. For CNN pansharpening, there are two major phases during the processing. In the first phase, the CNN model is solved with the training patches. In the second phase, the CNN model is used to pansharpen the MS image.

The learning properties of the compared CNN-based methods are summarized in Table I, which report our DiCNN1 and DiCNN2 built in pansharpening detail injection context. CNN-based pansharpening methods were trained using a GPU (Nvidia GTX 1060 3GB with CUDA 8.1 and CUDNN V5) through Caffe [47] in an Ubuntu 14.04 operating system, and tested on MATLAB R2016b via CPU mode (laptop with Intel I7 and 8GB RAM) through the deep learning framework Matconvnet [48] in Windows 10 operating system.

In addition to DiCNN1, DiCNN2, PNN, and DRPNN, several representative CS/MRA methods, including Gram Schmidt adaptive (GSA) [11], partial replacement adaptive component substitution (PRACS) [49], ATWT [16], band-dependent spatial-detail (BDSD) [50], and Generalized Laplacian pyramid with context-based decision (GLP-CBD) [2] are also tested for comparison. Notice that, there are no tricks, such as pansharpening phase fine-tuning and more shortcut connection, adopted in the considered CNN-based approaches, as the main purpose focuses at exploring the interpretabilities of detail injection based CNNs and the connected theoretical validation with some mathematical derivations on the simple forms as DiPNN1 and DiPNN2, respectively.

A. Experiment 1: WorldView-2 Washington Dataset

The dataset[1] was acquired with the WorldView-2 sensor over an urban area in Washington D.C., which provides a PAN image formed from wavelength 450–800 nm, and a MS image with eight bands, including four standard bands (blue, green, red, and near infrared 1) and four new bands (coastal, yellow, red edge, and near infrared 2). The resolution ratio R is 4, and the radiometric resolution is 11b, with the spatial resolution of the PAN image and that of the MS image being 0.46 m and 1.84 m, respectively. We choose two scenes with 256×256 pixels for tests in the reduced-resolution and full-resolution experiments separately. In the CNN model solving phase on this dataset, PNN and DiCNN1 take roughly 2.8 h, while DiCNN2 takes 2.5 h and DRPNN takes 3 h. This is due to the fact that the first convolutional layer involves fewer free parameters, whereas DRPNN contains an extra convolutional layer compared to other three CNNs.

[1] [Online]. Available: https://www.digitalglobe.com/resources/product-samples.

TABLE II
QUALITY INDEXES OF DIFFERENT PANSHARPENING METHODS UNDER A REDUCED-RESOLUTION QUALITY ASSESSMENT ON A 256×256 SUBSCENE OF A WORLDVIEW-2 DATASET

	Q8	SAM	ERGAS	SCC	Time(s)
Reference	1	0	0	1	
EXP	0.6726	7.9558	8.0358	0.5127	
GSA	0.9151	7.5830	4.3501	0.8973	0.85
PRACS	0.8682	7.7322	5.2648	0.8650	1.43
ATWT	0.8974	7.2241	4.7585	0.8926	**0.84**
BDSD	0.9178	8.1158	4.5293	0.8993	1.14
GLP-CBD	0.9148	7.5004	4.3438	0.8981	**0.84**
PNN	0.9243	7.6205	4.2924	0.8966	2.20
DRPNN	0.9325	7.2175	3.9664	0.9149	1.17
DiCNN1	**0.9492**	**6.2771**	**3.6487**	*0.9281*	1.13
DiCNN2	*0.9448*	*7.2012*	*3.7063*	**0.9299**	0.98

The best and second-best results are marked in bold and italic.

Table II shows the results of the reduced-resolution quality assessment. The computation times for the pansharpening phase are also included. As we can observe, CNN-based methods yield much better pansharpening quality than the CS-based and MRA-based methods. DiCNN1 and DiCNN2 achieve the highest Q8, SAM, ERGAS, and SCC scores among all compared methods, including CNN-based methods, and meanwhile DiCNN2 is the fastest among all CNN-based methods.

Fig. 4 displays the images of reduced-resolution experimental results. It shows that the pansharpened images yielded by CNN-based methods look much more similar to the ground-truth, without noticeable artifacts or spectral distortions. Fig. 5 shows the detail images, which are produced with the difference between the pansharpened HRMS image and the pre-interpolated LRMS image. The ground-truth details are achieved by the subtraction between the full-resolution MS image and the pre-interpolated one. The detail images are also in favor of the aforementioned observations, as it can be seen in the central circle area. For the CNN-based methods, the performances are hard to distinguish, but by investigating the spectral preservation of ground objects with small sizes, it is clear that DiCNN1 helps to impede spectral distortion more efficiently, as it can be seen in the bottom leftmost part of Fig. 4(h)–(k). Fig. 6 shows the residual images that are generated by the difference between the pansharpened HRMS image and the ground-truth image. From Fig. 6, we can see that the proposed methods exhibit very good performance.

Fig. 7 displays the full-resolution experimental results. The CNN-based methods exhibit sharper results than the other tested methods, especially in the vegetation areas. DiCNN1, PNN,

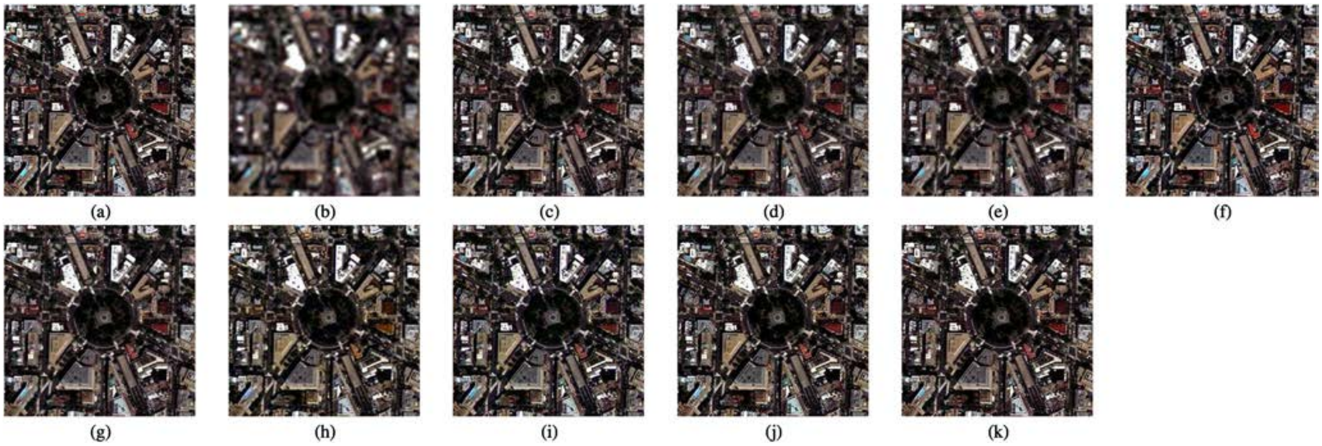


Fig. 4. Pansharpening results for a Worldview-2 dataset (composited with red, green, blue bands). (a) Ground-truth. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

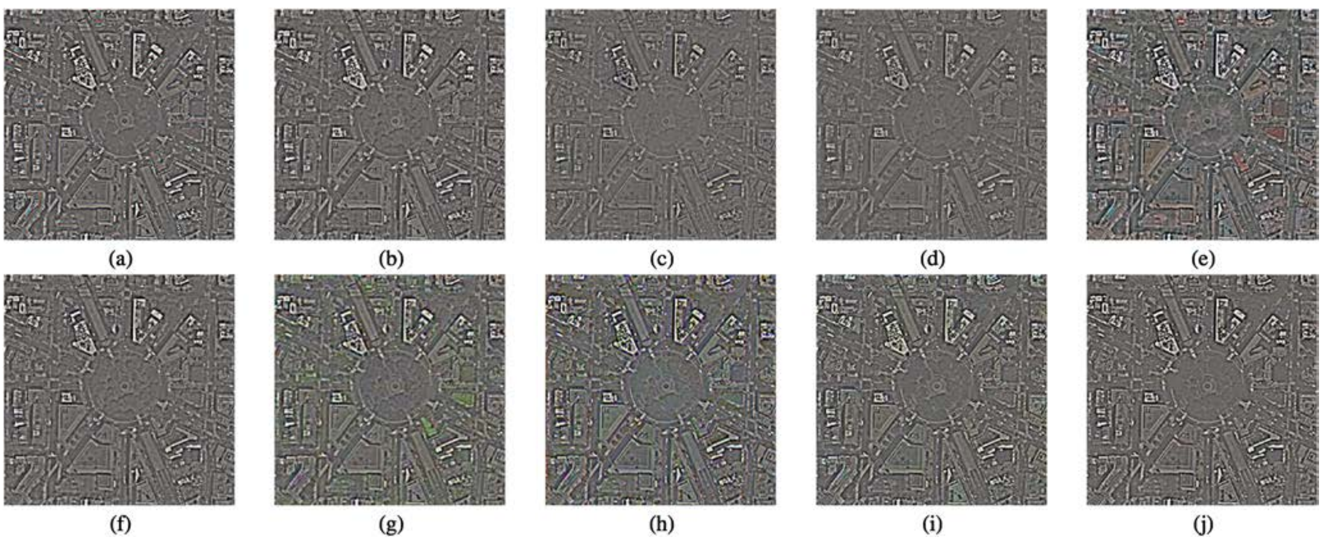


Fig. 5. Detail images of the Worldview-2 dataset. (a) Ground-truth. (b) GSA. (c) PRACS. (d) ATWT. (e) BDSD. (f) GLP-CBD. (g) PNN. (h) DRPNN. (i)DiCNN1. (j) DiCNN2.

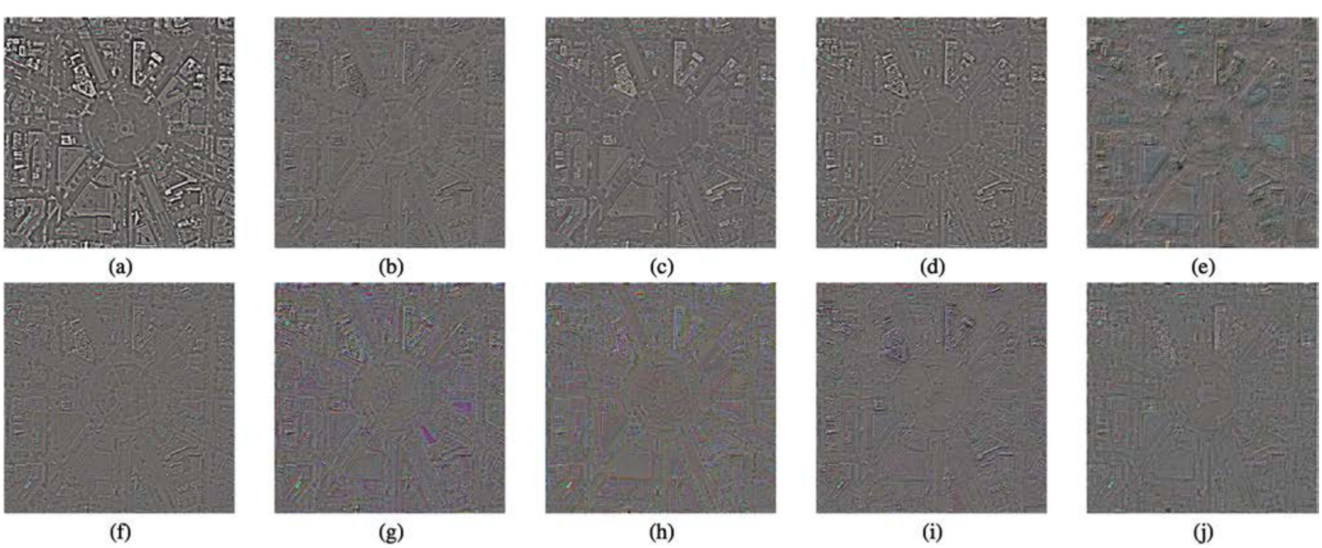


Fig. 6. Differences between the pansharpened images and the ground-truth of the Worldview-2 dataset. (a) EXP. (b) GSA. (c) PRACS. (d) ATWT. (e) BDSD. (f) GLP-CBD. (g) PNN. (h) DRPNN. (i)DiCNN1. (j) DiCNN2.

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

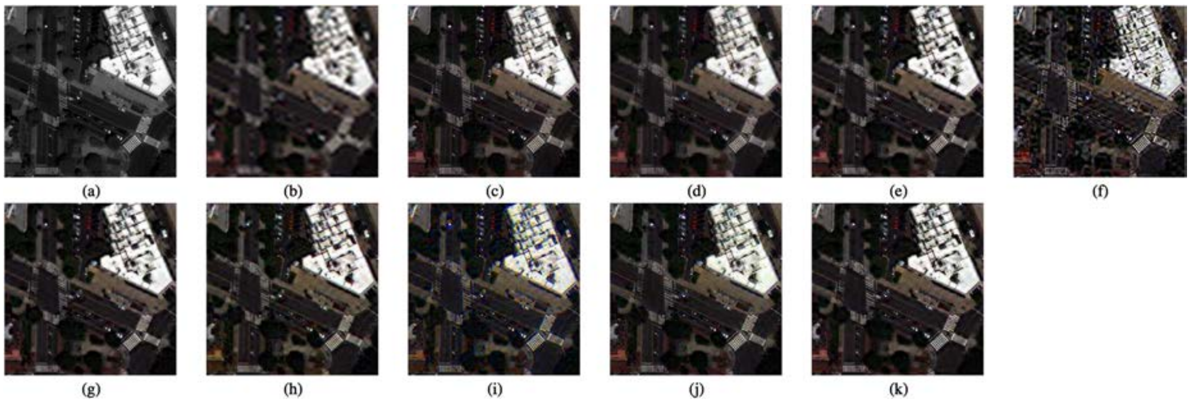


Fig. 7. Full-resolution pansharpening results for the WorldView-2 dataset. (a) PAN image. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

TABLE III
QUALITY INDEXES OF DIFFERENT PANSHARPENING METHODS UNDER REDUCED-RESOLUTION QUALITY ASSESSMENT ON A 256×256 SUBSCENE OF THE IKONOS DATASET

	Q4	SAM	ERGAS	SCC	Time(s)
Reference	1	0	0	1	
EXP	0.5791	5.4338	5.7489	0.5453	
GSA	0.8083	5.1063	4.1467	0.7583	0.73
PRACS	0.7843	5.1175	4.2096	0.7646	0.61
ATWT	0.8036	5.1198	4.0957	0.7622	**0.49**
BDSD	0.8141	5.4020	4.2070	0.7583	1.15
GLP-CBD	0.8121	5.0884	4.0857	0.7591	*0.52*
PNN	0.8846	4.8722	3.1783	0.8836	2.44
DRPNN	0.8995	4.5546	2.9513	0.9018	1.19
DiCNN1	**0.9120**	**4.3359**	**2.8532**	**0.9091**	1.38
DiCNN2	*0.9003*	*4.4575*	*2.9104*	*0.9027*	1.06

The best and second-best results are marked in bold and italic.

and DiCNN2 slightly overpass DRPNN in terms of reducing artifacts.

B. Experiment 2: IKONOS Hobart Dataset

The dataset[2] represents an urban and harbor area of Hobart, Australia. It was acquired by the IKONOS sensor, which collects data in the visible and near-infrared spectrum ranges. The MS sensor is characterized by four bands (blue, green, red, and near infrared) and also a PAN channel with band range from 450 to 900 nm. The resolution of MS is 4 m and PAN is 1 m. The radiometric resolution is 11b. Different areas with size of 256×256 pixels are used for reduced-resolution and full-resolution experiments, respectively.

Table III tabulates the results of our reduced-resolution quality assessment on the IKONOS Hobart dataset. Similar phenomena to the ones observed with the previous WorldView-2 dataset can be appreciated. Specifically, CNN-based methods achieve better pansharpening quality than the CS-based and MRA-based methods. DiCNN1 achieves the highest Q4, SAM, ERGAS, and SCC scores, while PNN is the most time consuming. DiCNN2 achieves the least computational time among CNN-based methods.

Fig. 8 displays the reduced-resolution experimental results. As it can be observed, CS/MRA-based methods exhibit poorer pansharpening results than CNN-based methods, as it can be seen in the edges of roofs shown in Fig. 8(c)–(g). Furthermore, DiCNN1 and DiCNN2 look most similar to the ground-truth in terms of spectral fidelity, as it can be seen in the vegetation area in the top left part of Fig. 8(j) and (k). Fig. 9 shows the detail images learned from various methods. They also support the previous observations and, additionally, confirm that DiCNN1 performs slightly better than DiCNN2 in terms of edge restoration, as it can be seen in the circle vegetation area in the bottom leftmost part of Fig. 9(i) and (j). Fig. 10 displays the full-resolution experimental results on IKONOS Hobart dataset. Similar observations can be made with regards to the experimental results reported for the WorldView-2 Washington dataset.

C. Experiment 3: Quickbird Sundarbans Dataset

The dataset[3] represents a forest area of Sundarbans in India. It was obtained by the QuickBird sensor, which provides a high-resolution PAN image with spectral cover range from 760 to 850 nm and with resolution of 0.6 m, and a four-band (blue, green, red and near infrared) MS image with resolution of 2.4 m. The radiometric resolution is also 11b. We selected different areas with size of 256×256 pixels for our reduced-resolution and full-resolution experiments, respectively.

Table IV shows the reduced-resolution quality assessment on the Chilika Lake dataset. We can easily conclude that similar phenomena also arise in this dataset. CNN-based methods achieve better pansharpening quality than CS-based and MRA-based methods. DiCNN1 overpasses others in terms of Q4, SAM, ERGAS, and SCC scores. DiCNN2 is the fastest among CNN-based methods, with comparable performance to DRPNN.

Fig. 11 displays the reduced-resolution experimental results. DiCNN1, DiCNN2, and DRPNN look much more similar to the original MS image, but DiCNN2 exhibits less ringing artifacts, such as the edges of the lakes in the leftmost part of Fig. 11(i)–(k). This phenomenon occurs more frequently in PNN. Meanwhile, EXP and PRACS result in significant

[2][Online]. Available: http://www.isprs.org/data/default.aspx.

[3][Online]. Available: http://glcf.umd.edu/data/quickbird/datamaps.shtml.

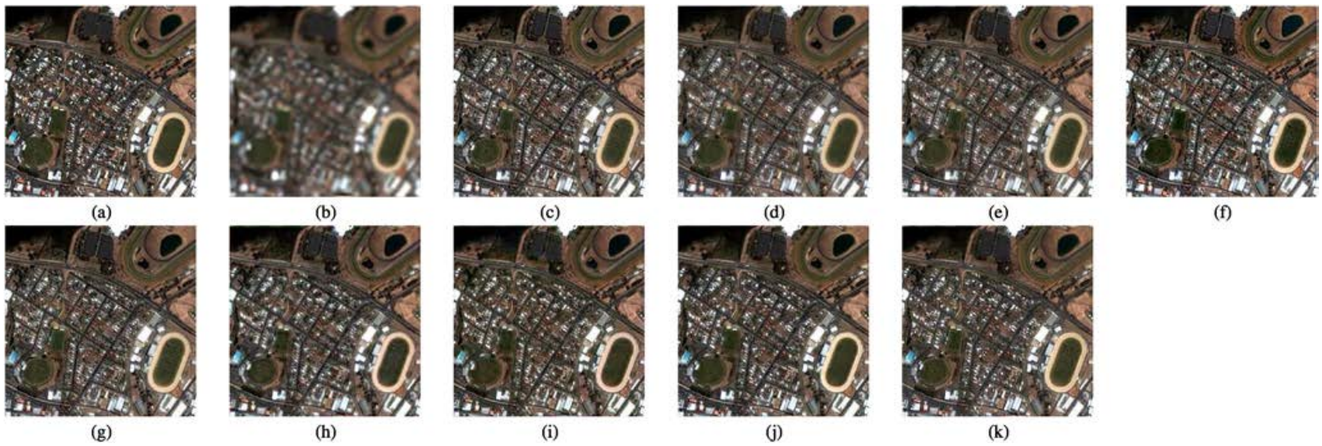


Fig. 8. Pansharpening results for IKONOS dataset. (a) Ground-truth. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

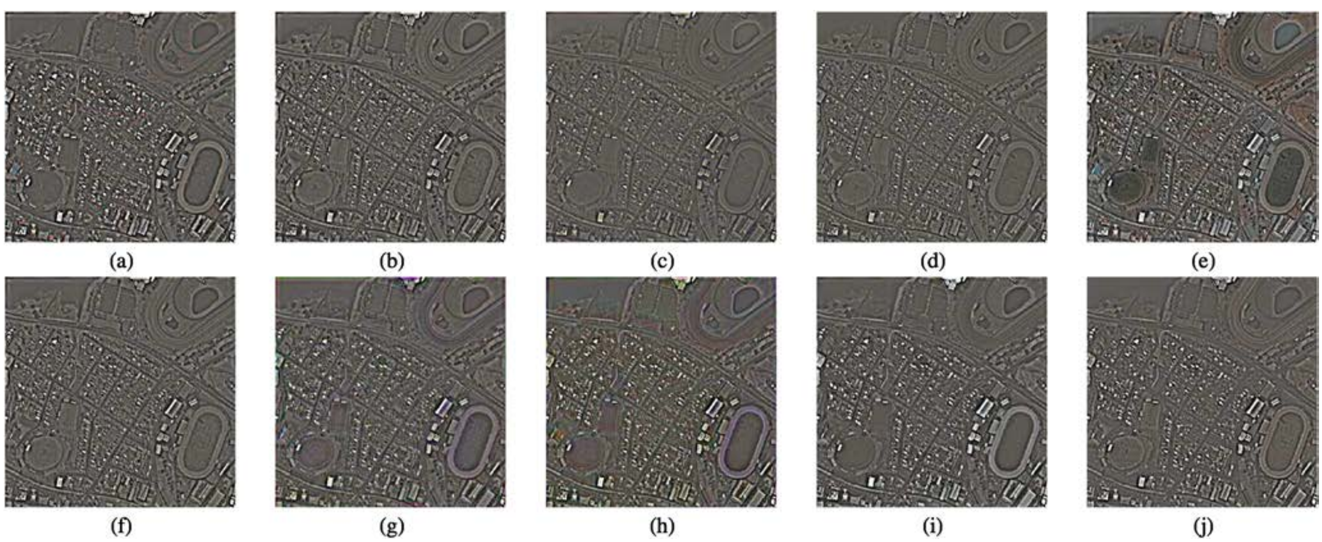


Fig. 9. Detail images of IKONOS dataset. (a) Ground-truth. (b) GSA. (c) PRACS. (d) ATWT. (e) BDSD. (f) GLP-CBD. (g) PNN. (h) DRPNN. (i) DiCNN1. (j) DiCNN2.

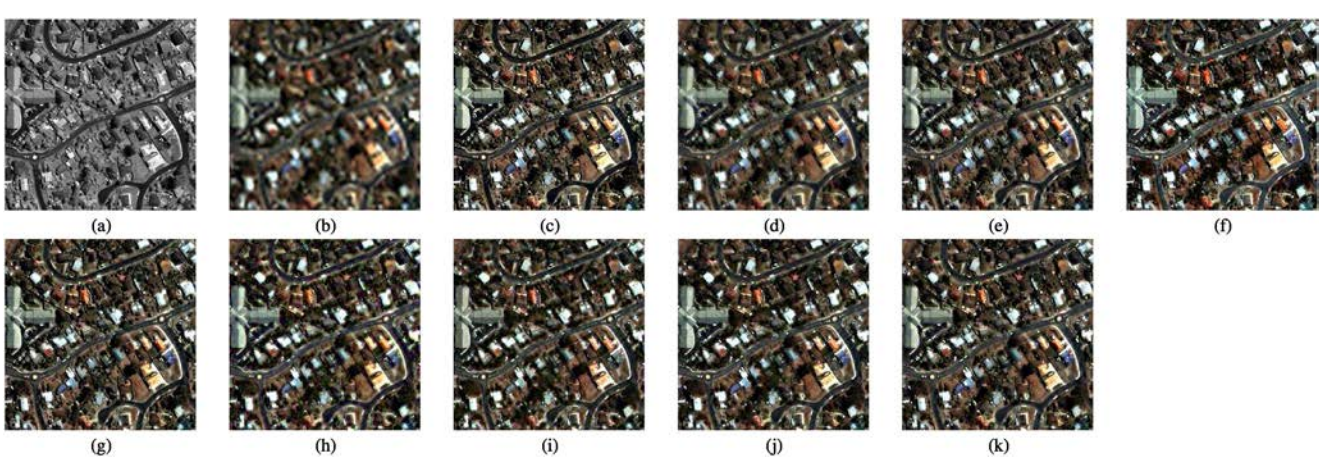


Fig. 10. Full-resolution pansharpening results for the IKONOS dataset. (a) PAN image. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

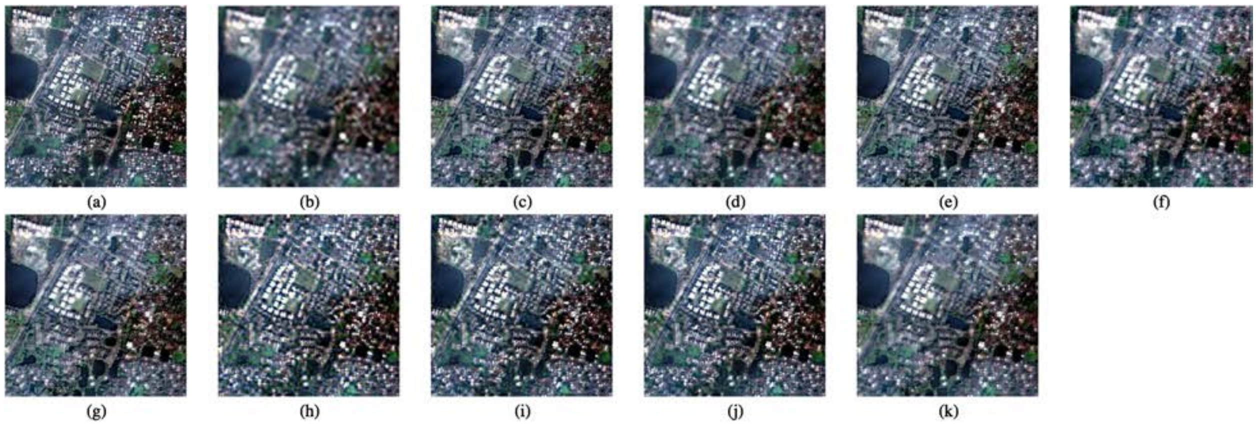


Fig. 11. Pansharpening results for the Quickbird dataset. (a) Ground-truth. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

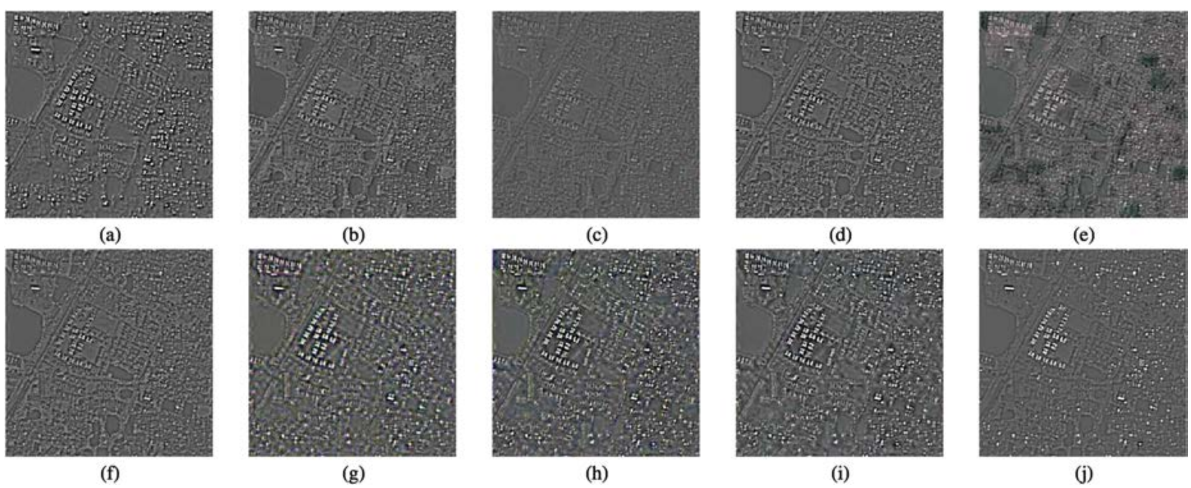


Fig. 12. Detail images of the Quickbird dataset. (a) Ground-truth. (b) GSA. (c) PRACS. (d) ATWT. (e) BDSD. (f) GLP-CBD. (g) PNN. (h) DRPNN. (i) DiCNN1. (j) DiCNN2.

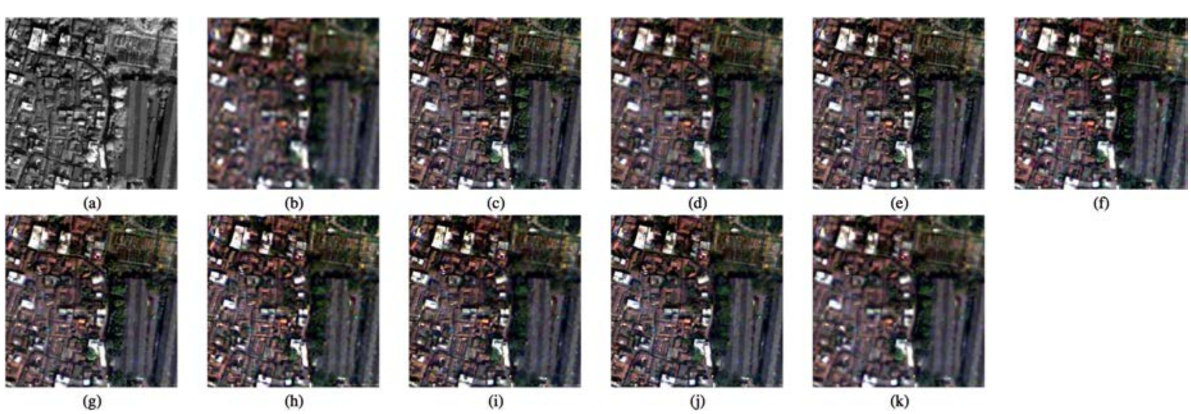


Fig. 13. Full-resolution pansharpening results for the Quickbird dataset. (a) PAN image. (b) EXP. (c) GSA. (d) PRACS. (e) ATWT. (f) BDSD. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) DiCNN2.

blurring effects. Fig. 12 shows the detail images learned from various methods, which also support the observations mentioned above. Fig. 13 displays the full-resolution experimental results. EXP has apparent blurring effects, whereas PRACS and DiCNN2 lead to subtle blurring effects. Blurring yielded from DiCNN2 is due to the fact that there is only the PAN image as the input of the convolution layers pathway, which makes it possible that the details complement to the LRMS image are learned insufficiently. In the meantime, DiCNN1 exhibits less artifacts than DRPNN and PNN.

TABLE IV
QUALITY INDEXES OF DIFFERENT PANSHARPENING METHODS UNDER REDUCED-RESOLUTION QUALITY ASSESSMENT ON A 256×256 SUBSCENE OF THE QUICKBIRD DATASET

	Q4	SAM	ERGAS	SCC	Time(s)
Reference	1	0	0	1	
EXP	0.6521	3.6555	3.0620	0.6615	
GSA	0.8321	3.4710	2.4565	0.8485	**0.13**
PRACS	0.7941	3.0063	2.2323	0.8501	*0.20*
ATWT	0.8361	2.9223	2.1011	0.8699	0.29
BDSD	0.8273	3.8008	2.6260	0.8378	0.15
MTF-GLP-CBD	0.8273	3.5584	2.5339	0.8488	0.41
PNN	0.8513	3.2265	2.0905	0.9153	0.31
DRPNN	*0.8979*	*2.5153*	*1.6278*	*0.9458*	0.37
DiCNN1	**0.9023**	**2.4674**	**1.6062**	**0.9464**	0.32
DiCNN2	0.8763	2.7850	1.7955	0.9317	0.22

The best and second-best results are marked in bold and italic.

It is worth noting that the spectral range of the PAN image in the Quickbird Sundarbans dataset spans only 90 nm wide, which is far narrower than that of the PAN images in WorldView-2 Washington dataset and IKONOS Hobart dataset, i.e., 350 and 402 nm. This implies that, when we perform pansharpening on the Quickbird Sundarbans Dataset, the PAN image will offer far less information to compensate for the spectral range difference between the PAN image and the MS image and mine the useful details. As we have mentioned previously, DiCNN1 and DiCNN2 have different structures. In DiCNN1, both the PAN image and the LRMS image are forwarded to the convolutional pathway, whereas in DiCNN2 only the PAN image is fed into the convolutional pathway, which means that DiCNN1 comprises two sources of information in its convolutional pathway and, hence, it exhibits higher potential to acquire information that is complementary to the low-resolution MS image. In contrast, DiCNN2 heavily relies on the PAN image to learn useful details for pansharpening. As discussed above, compared to DiCNN1, DiCNN2 depends much more on the PAN image to mine details for pansharpening. On the other hand, the PAN image in Quickbird Sundarbans Dataset has far narrower spectral range and, thus, it tends to offer far less information that is useful for pansharpening. Therefore, when both DiCNN1 and DiCNN2 are applied to Quickbird Sundarbans Dataset, DiCNN2 tends to yield worse pansharpening results than those achieved by DiCNN1, as shown in Fig. 11(j) and (k), and Fig. 13(j) and (k).

D. Experiment 4: Transfer Learning

To demonstrate the robustness of DiCNN2 under the situation that the number of bands of the test MS image has varied, we use the WorldView-2 Washington dataset and IKONOS Hobart Dataset in this experiment. Here, DiCNN2 is first trained on the original dataset. Then some of the MS bands are removed, and the final convolutional layers are fine-tuned to accommodate the current number of bands with 1.0×10^4 training iterations, much less than that in the previous training step. For the WorldView-2 Washington dataset with eight MS bands, four bands are removed. For the IKONOS Hobart Dataset with four MS bands, one band is removed.

Table V shows a quantitative assessment result on the WorldView-2 Washington dataset. As shown in the table, DiCNN2 yields the best scores in all evaluation metrics. It is remarkable that the time DiCNN2 needs for the training phase is less than half of the longest one, which results from the fact that DiCNN2 only needs to fine-tune the final convolutional layer.

TABLE V
QUALITY INDEXES OF CNN-BASED METHODS ON A 256×256 SUBSCENE OF A FOUR-BAND WORLDVIEW-2 DATASET

	Q4	SAM	ERGAS	SCC	Time(s)
Reference	1	0	0	1	
PNN	0.9308	3.4808	2.5678	0.9343	*222*
DRPNN	0.9462	3.0384	2.4160	0.9383	360
DiCNN1	*0.9497*	*2.8630*	*2.3080*	*0.9407*	327
DiCNN2	**0.9499**	**2.7148**	**2.2853**	**0.9420**	**173**

The best and second-best results are marked in bold and italic.

We also apply a similar experiment using the IKONOS data. Since the IKONOS dataset consists of four bands, we randomly choose three of them for testing. The four-band dataset is used to train DiCNN2, while the three-band one is applied to fine-tune the last layer of DiCNN2 and train other CNN-based methods.

Table VI tabulates the pansharpening results obtained by different CNN-based methods. As it can be observed, DiCNN2 outperforms others in most quality indexes. In addition, although DiCNN1 attains comparative results with regards to DiCNN2, the training time of the latter is far less than the former.

TABLE VI
QUALITY INDEXES OF CNN-BASED METHODS ON A 256×256 SUBSCENE OF THE THREE-BAND IKONOS DATASET

	Q4	SAM	ERGAS	SCC	Time(s)
Reference	1	0	0	1	
PNN	0.8748	2.4828	3.0206	0.8988	*176*
DRPNN	0.8928	2.8445	3.0683	0.9093	355
DiCNN1	**0.8989**	**1.9336**	*2.7109*	*0.9144*	335
DiCNN2	*0.8986*	*2.0317*	**2.6908**	**0.9181**	**160**

The best and second-best results are marked in bold and italic.

VI. CONCLUSION AND FUTURE LINES

In this paper, we have developed two CNN-based pansharpening methods, i.e., DiCNN1 and DiCNN2, based on a detail injection framework (DiPAN), which classical CS/MRA-base pansharpening methods can be ascribed into. In our newly developed DiCNN1 and DiCNN2, the MS details are learned in an end-to-end manner, which has explicit physical meaning and avoids separately dealing with injection gains and PAN details, as it is the case in traditional CS and MRA methods. Our DiCNN1 and DiCNN2 methods can gain low initial loss, which tends to yield faster convergence and exhibit excellent pansharpening performance. Particularly, DiCNN2 can additionally realize transfer learning when the type of the MS image or the PAN image changes, which is a highly desirable property. In the future, we will explore the possibility of designing PNNs with more hidden layers and more complex inter-connections among multiple convolutional layers.

APPENDIX

In this appendix, we provide a careful analysis of the efficiency of the proposed approach. The appendix includes two parts. First, we show that the proposed framework has good initialization. Then, we show that the proposed framework has good optimization.

A. Analysis of the Initialization of the Framework

CNN models are usually formulated as non-convex optimization problems with many local minima [51]–[53]. To solve such optimizations, the iterative gradient descent method is widely used, where the initialization and the gradient are usually critical for the solution.

Intuitively, better initializations are beneficial to attain better gradient descent solutions. Let us investigate such an initialization issue in more detail. For the four PNNs illustrated in Fig. 3, the output of the stacked convolutional layers pathway can be formulated as follows:

$$\mathbf{Z}_3 = \mathbf{W}_3 * \varphi(\mathbf{W}_2 * \varphi(\mathbf{W}_1 * \mathbf{X} + \mathbf{B}_1) + \mathbf{B}_2) + \mathbf{B}_3 \quad (13)$$

where $*$ denotes convolution, $\varphi(\cdot)$ represents the ReLU activation function, and $\mathbf{Z}_l = \mathbf{W}_l * \varphi(\mathbf{W}_{l-1} * \varphi(\mathbf{Z}_{l-1})$ denotes the output of the lth convolutional layer. $\mathbf{Z}_l$ is in 3-D data arrangement and thus a three-way tensor, the concept that has been previously mentioned in the description of (8). Note that $\mathbf{Z}_3$ has specific meanings for different PNNs, where it represents the MS details $\widehat{\mathbf{D}}$ for our DiCNN1 and DiCNN2, the residuals $\widehat{\mathbf{R}}$ for DRPNN, and the pansharpened HRMS image $\widehat{\mathbf{M}}$ for PNN.

In this paper, the initialization of CNN parameters $\mathbf{W}_l$ and $\mathbf{B}_l$ are assumed to follow an i.i.d. zero-mean random distribution and be independent of the neuron output of the $l-1$th layer $\mathbf{A}_{l-1} = \varphi(\mathbf{W}_{l-1} * \mathbf{A}_{l-2} + \mathbf{B}_{l-1})$. Obviously, the CNN input $\mathbf{X}$ can be used as $\mathbf{A}_0$. For later use, we present a property about $\mathbf{Z}_3$ and its proof below as

$$\begin{aligned}
&E\{\{\mathbf{Z}_3\}_{(1)}\mathbf{Y}\}\\
&\quad= E\{\{\{\mathbf{W}_3 * \varphi(\mathbf{Z}_2)\}_{(1)} + \{\mathbf{B}_3\}_{(1)}\}\mathbf{Y}\}\\
&\quad= E\{\{\mathbf{W}_3 * \varphi(\mathbf{Z}_2)\}_{(1)}\mathbf{Y}\} + E\{\{\mathbf{B}_3\}_{(1)}\mathbf{Y}\}\\
&\quad= E\left\{\left\{\sum_m\sum_n\sum_l \mathbf{W}_3(m,n,l)\right.\right.\\
&\qquad\left.\left.\times\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\right\}_{(1)}\mathbf{Y}\right\} + E\{\{\mathbf{B}_3\}_{(1)}\}E(\mathbf{Y})\\
&\quad= E\left\{\left\{\sum_m\sum_n\sum_l \mathbf{W}_3(m,n,l)\right.\right.\\
&\qquad\left.\left.\times\{\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\}_{(1)}\right\}\mathbf{Y}\right\} + \mathbf{0}\cdot E(\mathbf{Y})\\
&\quad= E\left\{\sum_m\sum_n\sum_l \mathbf{W}_3(m,n,l)\right.\\
&\qquad\left.\times\{\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\}_{(1)}\mathbf{Y}\right\}\\
&\quad= \sum_m\sum_n\sum_l\{0\cdot E\{\{\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\}_{(1)}\mathbf{Y}\}\}\\
&\quad= \mathbf{0}
\end{aligned} \quad (14)$$

where $\mathbf{Y}$ is a matrix not necessarily independent of $\mathbf{Z}_3$ and $\{\cdot\}_{(1)}$ means the unfolding of a three-way tensor along its first mode, and the following:

$$\begin{aligned}
&\{\mathbf{W}_3 * \varphi(\mathbf{Z}_2)\}_{(1)}\\
&\quad= \left\{\sum_m\sum_n\sum_l \mathbf{W}_3(m,n,l)\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\right\}_{(1)}\\
&\quad\times\sum_m\sum_n\sum_l \mathbf{W}_3(m,n,l)\\
&\quad\times\{\varphi(\mathbf{Z}_2(m-x,n-y,l-b))\}_{(1)}
\end{aligned} \quad (15)$$

are utilized.

We will first justify that our DiCNNs can achieve better initialization. First, consider DiCNN1. Its loss function $E(\|\widehat{\mathbf{D}} + \widetilde{\mathbf{M}} - \mathbf{Y}\|_F^2)$ can be rewritten as

$$\begin{aligned}
&E(\|\widehat{\mathbf{D}} + \widetilde{\mathbf{M}} - \mathbf{Y}\|_F^2)\\
&\quad= E\{\mathrm{Trace}\{(\widehat{\mathbf{D}} + \widetilde{\mathbf{M}} - \mathbf{Y})(\widehat{\mathbf{D}} + \widetilde{\mathbf{M}} - \mathbf{Y})^T\}\}\\
&\quad= E\{\mathrm{Trace}(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T) + \mathrm{Trace}(\widehat{\mathbf{D}}\widetilde{\mathbf{M}}^T) - \mathrm{Trace}(\widehat{\mathbf{D}}\mathbf{Y}^T)\\
&\qquad+ \mathrm{Trace}(\widetilde{\mathbf{M}}\widehat{\mathbf{D}}^T) + \mathrm{Trace}(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T) - \mathrm{Trace}(\widetilde{\mathbf{M}}\mathbf{Y}^T)\\
&\qquad- \mathrm{Trace}(\mathbf{Y}\widehat{\mathbf{D}}^T) - \mathrm{Trace}(\mathbf{Y}\widetilde{\mathbf{M}}^T)\\
&\qquad+ \mathrm{Trace}(\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= E\{\mathrm{Trace}(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T) + 2\mathrm{Trace}(\widehat{\mathbf{D}}\widetilde{\mathbf{M}}^T) - 2\mathrm{Trace}(\widehat{\mathbf{D}}\mathbf{Y}^T)\\
&\qquad+ \mathrm{Trace}(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T) - 2\mathrm{Trace}(\widetilde{\mathbf{M}}\mathbf{Y}^T)\\
&\qquad+ \mathrm{Trace}(\mathbf{Y}\mathbf{Y}^T)\}\\
&\qquad- 2\mathrm{Trace}\{E(\widehat{\mathbf{D}}\mathbf{Y}^T)\} + \mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}\\
&\qquad- 2\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\mathbf{Y}^T)\} + \mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T)\} + \mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}\\
&\qquad- 2\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\mathbf{Y}^T)\} + \mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}
\end{aligned} \quad (16)$$

where the equations

$$\mathrm{Trace}\{E(\widehat{\mathbf{D}}\widetilde{\mathbf{M}}^T)\} = \mathbf{0} \quad (17)$$

$$\mathrm{Trace}\{E(\widehat{\mathbf{D}}\mathbf{Y}^T)\} = \mathbf{0} \quad (18)$$

are utilized, which can be obtained through (14).

Let us now consider PNN; its loss function $E(\|\widehat{\mathbf{M}}-\mathbf{Y}\|_F^2)$ can be transformed as

$$\begin{aligned}
&E(\|\widehat{\mathbf{M}}-\mathbf{Y}\|_F^2)\\
&\quad= E\{\mathrm{Trace}\{(\widehat{\mathbf{M}}-\mathbf{Y})(\widehat{\mathbf{M}}-\mathbf{Y})^T\}\}\\
&\quad= E\{\mathrm{Trace}(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T-\widehat{\mathbf{M}}\mathbf{Y}^T-\mathbf{Y}\widehat{\mathbf{M}}^T+\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= E\{\mathrm{Trace}(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)-2\mathrm{Trace}(\widehat{\mathbf{M}}\mathbf{Y}^T)+\mathrm{Trace}(\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)\}-2\mathrm{Trace}\{E(\widehat{\mathbf{M}}\mathbf{Y}^T)\}\\
&\qquad+\mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)\}+\mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}
\end{aligned}\tag{19}$$

where the equation

$$\mathrm{Trace}\{E(\widehat{\mathbf{M}}\mathbf{Y}^T)\}=\mathbf{0}\tag{20}$$

is involved, which can also be obtained via (14).

Recall that $\widetilde{\mathbf{M}}$ represents the pre-interpolated LRMS and $\mathbf{Y}$ denotes the ideal HRMS. Therefore, $(\widetilde{\mathbf{M}}-\mathbf{Y})$ represent MS details whose energy tends to be significantly less than that of pre-interpolated LRMS. To compare the initialization of loss function of DiCNN1 shown in (16) with that of PNN shown in (19), we have

$$\begin{aligned}
&E(\|\widehat{\mathbf{D}}+\widetilde{\mathbf{M}}-\mathbf{Y}\|_F^2)-E(\|\widehat{\mathbf{M}}-\mathbf{Y}\|_F^2)\\
&\quad= \mathrm{Trace}\{E(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T)\}+\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}\\
&\qquad-2\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\mathbf{Y}^T)\}+\mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}\\
&\qquad-\mathrm{Trace}\{E(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)\}-\mathrm{Trace}\{E(\mathbf{Y}\mathbf{Y}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T)\}+\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}\\
&\qquad-2\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\mathbf{Y}^T)\}-\mathrm{Trace}\{E(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}-2\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\mathbf{Y}^T)\}\\
&\quad= \mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}-2\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\mathbf{Y}^T+\widetilde{\mathbf{M}}^T-\widetilde{\mathbf{M}}^T)\}\}\\
&\quad= -\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}-2\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\mathbf{Y}^T-\widetilde{\mathbf{M}}^T)\}\}\\
&\quad= 2\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\widetilde{\mathbf{M}}^T-\mathbf{Y}^T)\}\}-\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}\\
&\quad< 0
\end{aligned}\tag{21}$$

where the equation

$$\mathrm{Trace}\{E(\widehat{\mathbf{D}}\widehat{\mathbf{D}}^T)\}=\mathrm{Trace}\{E(\widehat{\mathbf{M}}\widehat{\mathbf{M}}^T)\}\tag{22}$$

is utilized during the derivation from step 2 to step 3. This is reasonable, as $\widehat{\mathbf{D}}$ and $\widehat{\mathbf{M}}$ stand for the outputs of convolutional layers pathways of DiCNN1 and PNN, respectively. In the initial phases of these two CNNs, their convolutional layers pathways have similar structure, similar inputs, and the same distributed network parameters. Moreover, the diagonal entries of $\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T$ are always greater than or equal to zero. But, in a real image scenario, it is impossible that all of the diagonal entries are equal to zero. Accordingly, we have

$$\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}>0.\tag{23}$$

Taking a close inspection of the term $2\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\widetilde{\mathbf{M}}^T-\mathbf{Y}^T)\}\}$ in the last equality of (21), we find that $(\widetilde{\mathbf{M}}^T-\mathbf{Y}^T)$ exactly represents the ideal MS details whose energy should account for small portion that of the HRMS image and, thus, we have

$$\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}>2|\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\widetilde{\mathbf{M}}^T-\mathbf{Y}^T)\}\}|.\tag{24}$$

After using (23) and (24), problem (21) results in

$$E(\|\widehat{\mathbf{D}}+\widetilde{\mathbf{M}}-\mathbf{Y}\|_F^2)<E(\|\widehat{\mathbf{M}}-\mathbf{Y}\|_F^2).\tag{25}$$

In summary, we can conduct that, the initial loss of DiCNN1 is smaller than that of PNN. For verification, let

$$T_1=\mathrm{Trace}\{E(\widetilde{\mathbf{M}}\widetilde{\mathbf{M}}^T)\}$$

and

$$T_2=2|\mathrm{Trace}\{E\{\widetilde{\mathbf{M}}(\widetilde{\mathbf{M}}^T-\mathbf{Y}^T)\}\}|$$

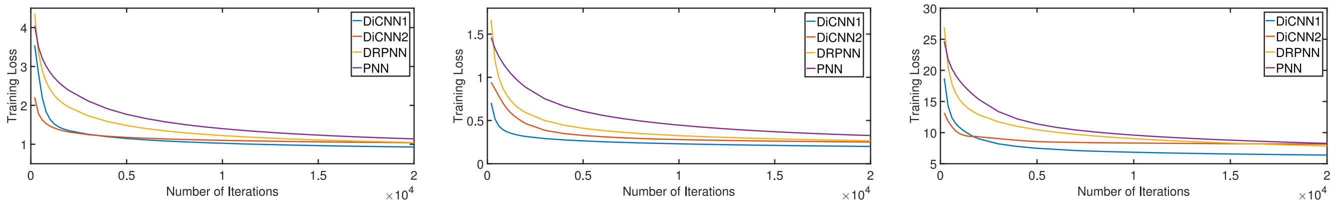
be the two traces in (24); Table VII illustrates T_1 and T_2 computed on three real datasets, which clearly shows that $T_1 \gg T2$, verifying that DiCNN1 has a better initialization than PNN.

TABLE VII
TRACE VALUES

	T_1	T_2
IKONOS	203.8785	2.9
Quickbird	108.138	1.1619
Worldview-2	607.1628	20.2275

B. Analysis of the Optimization of the Framework

Since the representations for the gradients are sophisticated and incomparable, it is difficult to quantitatively assess the influence of the gradients on the optimization processes of the four PNNs. Here, we resort to an empirical analysis instead. Fig. 14 illustrates the training losses of the four CNN methods on three datasets. It is observable that the initial losses of DiCNN1 and DiCNN2 are less than those of PNN and DRPNN, corresponding to the theoretical analysis presented earlier in the Appendix A, which means that DiCNN1 and DiCNN2 can achieve better initializations. PNN not only exhibits worse initialization, but also its iteration process (involving its gradient) does not change the inferior tendency of its loss. During the iterative process, PNN always yields a loss higher than that of DiCNN1 and DiCNN2. That is, the impact of the gradient-based iteration process is not strong enough to compensate for the loss resulting from an inappropriate initialization. DRPNN exhibits the worst initialization. Although its gradient-involved iterative process makes its loss drop fast, it is still always higher than that of DiCNN1 during the iterative process.



(a) (b) (c)

Fig. 14. Training losses of DiCNN1, DiCNN2, PNN, and DRPNN. (a) IKONOS image. (b) Quickbird image. (c) WorldView-2 image.

References

[1] G. Vivone *et al.*, "A critical comparison among pansharpening algorithms," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2565–2586, May 2015.

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

[3] L. Loncan *et al.*, "Hyperspectral pansharpening: A review," *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 27–46, Sep. 2015.

[4] 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.

[5] P. S. Chavez Jr., and A. Y. Kwarteng, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogrammetric Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[6] V. K. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogrammetric Eng. Remote Sens.*, vol. 58, no. 5, pp. 561–567, 1992.

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

[8] A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images channel ratio and chromaticity transformation techniques," *Remote Sens. Environ.*, vol. 22, no. 3, pp. 343–365, 1987.

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

[10] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6,011,875, Jan. 2000.

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

[12] M. M. Khan, J. Chanussot, L. Condat, and A. Montanvert, "Indusion: Fusion of multispectral and panchromatic images using the induction scaling technique," *IEEE Geosci. Remote Sens. Lett.*, vol. 5, no. 1, pp. 98–102, Jan. 2008.

[13] G. P. Nason and B. W. Silverman, "The stationary wavelet transform and some statistical applications," *Science*, vol. 346, no. 6212, pp. 918–299, 1995.

[14] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogrammetric Eng. Remote Sens.*, vol. 66, no. 2, pp. 49–61, 2000.

[15] 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.

[16] G. Vivone, R. Restaino, M. D. Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 5, pp. 930–934, May 2014.

[17] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "An MTF-based spectral distortion minimizing model for pan-sharpening of very high resolution multispectral images of urban areas," in *Proc. Workshop Remote Sens. Data Fusion Over Urban Areas*, 2003, pp. 90–94.

[18] B. Aiazzi, L. Alparone, S. Baronti, A. Garzelli, and M. Selva, "MTF-tailored multiscale fusion of high-resolution MS and Pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.

[19] J. Lee and C. Lee, "Fast and efficient panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 1, pp. 155–163, Jan. 2010.

[20] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, pp. 594–615, 2016.

[21] C. Thomas, T. Ranchin, L. Wald, and J. Chanussot, "Synthesis of multispectral images to high spatial resolution: A critical review of fusion methods based on remote sensing physics," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1301–1312, May 2008.

[22] W. Liao *et al.*, "Processing of multiresolution thermal hyperspectral and digital color data: Outcome of the 2014 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 8, no. 6, pp. 2984–2996, Jun. 2015.

[23] C. Dong, C. L. Chen, 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.

[24] C. Dong, C. L. Chen, K. He, and X. Tang, *Learning a Deep Convolutional Network for Image Super-Resolution*. New York, NY, USA: Springer, 2014.

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

[26] 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.*, 2016, pp. 1646–1654.

[27] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.

[28] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-sharpening," in *Proc. Int. Workshop Remote Sens. With Intell. Process*, 2017, pp. 1–4.

[29] K. Fukushima, "Neocognitron: A hierarchical neural network capable of visual pattern recognition," *Neural Netw.*, vol. 1, no. 2, pp. 119–130, 1988.

[30] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[33] W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *J. Sens.*, vol. 2015, 2015, Art. no. 258619.

[34] 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.

[35] X. Jia, "Single image super-resolution using multi-scale convolutional neural network," in *Proc. Pacific-Rim Conf. Multimedia*, 2017, vol. 8, no. 7, pp. 149–157.

[36] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM Rev.*, vol. 51, no. 3, pp. 455–500, Aug. 2009.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 770–778.

[38] Q. Wei, "Bayesian fusion of multi-band images: A powerful tool for super-resolution," Ph.D. dissertation, Toulouse Inst. Tech., Toulouse, France, 2015.

[39] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
[40] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogrammetric Eng. Remote Sens.*, vol. 63, no. 6, pp. 691–699, 1997.
[41] L. Alparone and B. Aiazzi, "MTF-tailored multiscale fusion of high-resolution ms and pan imagery," *Photogrammetric Eng. Remote Sens.*, vol. 72, no. 5, pp. 591–596, 2006.
[42] L. Alparone, S. Baronti, A. Garzelli, and F. Nencini, "A global quality measurement of pan-sharpened multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 313–317, Oct. 2004.
[43] J. Zhou, D. L. Civco, and J. A. Silander, "A wavelet transform method to merge Landsat TM and SPOT panchromatic data," *Int. J. Remote Sens.*, vol. 19, no. 4, pp. 743–757, 1998.
[44] R. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Descrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," Tech. Rep. 19940012238, Dec. 1995. [Online]. Available: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19940012238.pdf
[45] L. Wald, "Data fusion. Definitions and architectures—Fusion of images of different spatial resolutions," Paris, France: Presses des MINES, 2002, ch. 8.
[46] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representa.*, 2015, pp. 1–15.
[47] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.
[48] A. Vedaldi and K. Lenc, "Matconvnet: Convolutional neural networks for MATLAB," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[49] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2010.
[50] 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. 2007.
[51] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, and Y. LeCun, "The loss surfaces of multilayer networks," in *Proc. Artif. Intell. Statist.*, 2015, pp. 192–204.
[52] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.
[53] S. S. Haykin, S. S. Haykin, S. S. Haykin, and S. S. Haykin, *Neural Networks and Learning Machines*, vol. 3. Upper Saddle River, NJ, USA: Pearson, 2009.

Lin He (S'05–M'12) received the B.S. degree from the Xi'an Institute of Technology, Xi'an, China, in 1995, the M.S. degree from Chongqing University, Chongqing, China, in 2003, both in instrumentation engineering, and the Ph.D. degree in pattern recognition and intelligent systems from Northwestern Polytechnical University, Xi'an, in 2007.

Since 2007, he has been with the School of Automation Science and Engineering, South China University of Technology, Guangzhou, China, where he is currently an Associate Professor. His current research interests include statistical pattern recognition, hyperspectral image processing, and high-dimensional signal processing.

Yizhou Rao received the B.E. degree in automation from the Guangdong University of Technology, Guangzhou, China, in 2015, and the M.S. degree in pattern recognition and intelligent system from the South China University of Technology, Guangzhou, in 2018.

His research interests include pansharpening, image fusion, and signal processing.

Jun Li (SM'16) received the Geographical Information Systems degree from Hunan Normal University, Changsha, China, in 2004, the M.Sc. degree in remote sensing and photogrammetry from Peking University, Beijing, China, in 2007, and the Ph.D. degree in electrical and computer engineering from the Instituto Superior Tecnico, Technical University of Lisbon, Lisbon, Portugal, in 2011.

From 2011 to 2012, she was a Postdoc Researcher with the Department of Technology of Computers and Communications, University of Extremadura, Badajoz, Spain. From 2014 to 2018, she was a Professor with the School of Geography and Planning, Sun Yat-Sen University, China. She is currently with the College of Electrical and Information Engineering, Hunan University. Since then, she has received several prestigious funding grants at the national and international level. She has published a total of 69 journal citation report (JCR) papers, 48 conference international conference papers, and one book chapter. She has received a significant number of citations to her published works, with several papers distinguished as "Highly Cited Papers" in Thomson Reuters' Web of Science—Essential Science Indicators (WoS-ESI). Her main research interests include remotely sensed hyperspectral image analysis, signal processing, supervised/semisupervised learning, and active learning.

Dr. Li has served as a Guest Editor of a Special Issue in the prestigious journal of the PROCEEDINGS OF THE IEEE. She has also served as a Guest Editor of a Special Issue in the prestigious *ISPRS Journal of Photogrammetry and Remote Sensing*. She has been an Associate Editor for the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING since 2014. Her students have also obtained important distinctions and awards at international conferences and symposia.

Jocelyn Chanussot (M'04–SM'04–F'12) received the M.Sc. degree in electrical engineering from the Grenoble Institute of Technology (Grenoble INP), Grenoble, France, in 1995, and the Ph.D. degree from the Universit de Savoie, Annecy, France, in 1998.

In 1999, he was with the Geography Imagery Perception Laboratory for the Delegation Generale de l'Armement (DGA—French National Defense Department). Since 1999, he has been with Grenoble INP, where he is currently a Professor of signal and image processing. He is conducting his research at the Grenoble Images Speech Signals and Automatics Laboratory (GIPSA-Lab). He has been a Visiting Scholar with Stanford University (USA), KTH (Sweden), and NUS (Singapore). Since 2013, he is an Adjunct Professor with the University of Iceland, Reykjavk, Iceland. In 2015–2017, he was a Visiting Professor with the University of California, Los Angeles. His research interests include image analysis, multicomponent image processing, nonlinear filtering, data fusion, and machine learning in remote sensing.

Dr. Chanussot is a member of the IEEE Geoscience and Remote Sensing Society AdCom. He was the General Chair of the first IEEE GRSS Workshop on Hyperspectral Image and Signal Processing, Evolution in Remote Sensing (WHISPERS). He was the Chair (2009–2011) and Co-Chair of the GRS Data Fusion Technical Committee (2005–2008). He was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society (2006–2008) and the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing (2009). He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and IEEE TRANSACTIONS ON IMAGE PROCESSING. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (2011–2015). In 2013, he was a Guest Editor for the PROCEEDINGS OF THE IEEE and, in 2014, a Guest Editor for the IEEE SIGNAL PROCESSING MAGAZINE. He is a member of the Institut Universitaire de France (2012–2017) and a 2018 Highly Cited Researcher (Clarivate Analytics). He is the founding President of IEEE Geoscience and Remote Sensing French chapter (2007–2010), which received the 2010 IEEE GRS-S Chapter Excellence Award. He was the co-recipient of the NORSIG 2006 Best Student Paper Award, the IEEE GRSS 2011 and 2015 Symposium Best Paper Award, the IEEE GRSS 2012 Transactions Prize Paper Award, and the IEEE GRSS 2013 Highest Impact Paper Award.

Antonio Plaza (F'15) received the M.Sc. degree in 1999 and the Ph.D. degree in 2002 from the University of Extremadura, Badajoz, Spain, both in computer engineering.

He is the Head of the Hyperspectral Computing Laboratory, Department of Technology of Computers and Communications, University of Extremadura, Badajoz, Spain. He has authored more than 600 publications, including 200 JCR journal papers (145 in IEEE journals), 23 book chapters, and 285 peer-reviewed conference proceeding papers. He has guest edited ten special issues on hyperspectral remote sensing for different journals. He has reviewed more than 500 manuscripts for more than 50 different journals. His main research interests include hyperspectral data processing and parallel computing of remote sensing data.

Dr. Plaza was the Director of Education Activities for the IEEE Geoscience and Remote Sensing Society (GRSS) in 2011–2012, and the President of the Spanish Chapter of IEEE GRSS in 2012–2016. He is currently the Editor-in-Chief for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING JOURNAL. He is also an Associate Editor for the IEEE ACCESS, and was a member of the Editorial Board of the IEEE GEOSCIENCE AND REMOTE SENSING NEWSLETTER (2011–2012) and the IEEE GEOSCIENCE AND REMOTE SENSING MAGAZINE (2013). He was also a member of the Steering Committee of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS). He was a recipient of the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (in 2009) and a recipient of the recognition of Best Reviewers of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (in 2010), for which he served as an Associate Editor in 2007–2012. He was a recipient of the Best Column Award of the IEEE SIGNAL PROCESSING MAGAZINE in 2015, the 2013 Best Paper Award of the JSTARS journal, and the most highly cited paper (2005–2010) in the *Journal of Parallel and Distributed Computing*. He received the best paper awards at the IEEE International Conference on Space Technology and the IEEE Symposium on Signal Processing and Information Technology. He is a Fellow of the IEEE "for contributions to hyperspectral data processing and parallel computing of Earth observation data." For additional information contact http://www.umbc.edu/rssipl/people/aplaza.

Jiawei Zhu received the B.E. degree in automation from the South China University of Technology, Guangzhou, China, in 2018, where he is currently working toward the M.S. degree.

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

Bo Li received the B.S. degree in computer science from Chongqing University, Chongqing, China, in 1986, the M.S. degree in computer science from Xian Jiaotong University, Xi'an, China, in 1989, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 1993.

Since 1993, he joined the School of Computer Science and Engineering, Beihang University. He is currently the Director of the Beijing Key Laboratory of Digital Media, and the Director of the Professional Committee of Multimedia Technology of the China Computer Federation. He has published more than 100 academic papers in diverse research fields, including intelligent perception, data mining, remote sensing image fusion, and intelligent hardware.