Detail Injection-Based Deep Convolutional Neural Networks for Pansharpening

Liang-Jian Deng, *Member, IEEE*, Gemine Vivone, *Senior Member, IEEE*, Cheng Jin, and Jocelyn Chanussot, *Fellow, IEEE*

***Abstract*— The fusion of high spatial resolution panchromatic (PAN) data with simultaneously acquired multispectral (MS) data with the lower spatial resolution is a hot topic, which is often called pansharpening. In this article, we exploit the combination of machine learning techniques and fusion schemes introduced to address the pansharpening problem. In particular, deep convolutional neural networks (DCNNs) are proposed to solve this issue. The latter is combined first with the traditional component substitution and multiresolution analysis fusion schemes in order to estimate the nonlinear injection models that rule the combination of the upsampled low-resolution MS image with the extracted details exploiting the two philosophies. Furthermore, inspired by these two approaches, we also developed another DCNN for pansharpening. This is fed by the direct difference between the PAN image and the upsampled low-resolution MS image. Extensive experiments conducted both at reduced and full resolutions demonstrate that this latter convolutional neural network outperforms both the other detail injection-based proposals and several state-of-the-art pansharpening methods.**

***Index Terms*— Component substitution (CS), deep convolutional neural network (DCNN), image fusion, multiresolution analysis (MRA), pansharpening, remote sensing.**

I. Introduction

PANSHARPENING has become a fundamental problem in remote sensing image processing since it can fuse a textithigh spatial resolution panchromatic (PAN) image and a low spatial resolution multispectral (MS) image in order to obtain an MS image with the highest (PAN) spatial resolution. PAN and MS images are quite common in the field of remote sensing imaging, and they are usually simultaneously acquired by sensors mounted on many satellites, such as IKONOS, WorldView-2, and WorldView-3. Pansharpening has attracted the interest of the scientific community. This is justified by the contest launched by the Data Fusion Committee of the IEEE Geoscience and Remote Sensing Society in 2006 [3], [4] and many recently published review articles [5], [6]. Furthermore, pansharpened products have attracted the interest of some commercial companies, e.g., Google Earth, and pansharpening has been exploited as preliminary a step for several image processing tasks, e.g., change detection [7], [8].

Most of the pansharpening works can be divided into four categories, i.e., component substitution (CS) methods, multiresolution analysis (MRA) approaches, variational optimization-based (VO) techniques, and machine learning (ML) approaches.

CS and MRA approaches play an important role in the community of pansharpening. They have shown promising performance with a balanced computational burden. The CS-based methods rely on the concept of the projection of the MS image into a new domain, where the spatial information can be easily separated into a component, usually called the intensity component. Then, the (possibly equalized) PAN image can be substituted with the intensity component. The sharpened version of the MS image is obtained due to the inverse projection bringing the data to the original MS domain. CS-based methods can generate outcomes with high spatial fidelity paid by a usually greater spectral distortion. Some powerful instances of methods belonging to this category are the band-dependent spatial-detail (BDSD) with local parameter estimation [9], the robust (BDSD-PC) method [10], the partial replacement adaptive component substitution (PRACS) [11], and Gram–Schmidt (GS) spectral sharpening [12].

MRA-based approaches inject spatial details extracted from the PAN image through an MRA framework into the MS image in order to get the high spatial resolution MS image. MRA-based products preserve spectral information but can suffer from spatial distortion. The examples of methods into this class are the smoothing filter-based intensity modulation (SFIM) [13], the additive wavelet luminance proportional (AWLP) [14], the “à-trous” wavelet transform [15], the Laplacian pyramid (LP) [16], the generalized Laplacian pyramid (GLP) [17], [18], the GLP with robust regression [19], and the GLP with full-scale regression (GLP-Reg) [20].

Recently, VO approaches have shown competitive ability in addressing the pansharpening issue. Techniques belonging to this class include the Bayesian methods [21]–[23],

Manuscript received June 26, 2020; revised September 7, 2020; accepted October 12, 2020. Date of publication October 27, 2020; date of current version July 22, 2021. The work of Liang-Jian Deng was supported in part by the NSFC under Grant 61702083, Grant 61772003, and Grant 61876203. *(Corresponding author: Gemine Vivone.)*

Liang-Jian Deng is with the School of Mathematical Sciences, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: liangjian.deng@uestc.edu.cn).

Gemine Vivone is with the National Research Council – Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy (e-mail: gemine.vivone@imaa.cnr.it).

Cheng Jin is with the School of Optoelectronics, University of Electronic Science and Engineering of China, Chengdu 611731, China (e-mail: cheng.jin@std.uestc.edu.cn).

Jocelyn Chanussot is with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, Laboratoire Jean Kuntzmann (LJK), 38000 Grenoble, France (e-mail: jocelyn.chanussot@grenoble-inp.fr).

Digital Object Identifier 10.1109/TGRS.2020.3031366

variational approaches [24]–[36], and compressed sensing techniques [37]–[39]. Despite their formal mathematical elegance, VO approaches provide only incremental performance improvements with respect to the state-of-the-art of CS and MRA methods; such improvement comes at the cost of high computational burden and presence of many parameters to be tuned explaining why CS and MRA are nowadays commonly advocated both for benchmarking and practical uses.

With the tremendous improvements of hardware, convolutional neural networks (CNNs) have recently become a powerful tool to deal with pansharpening and its related applications (see [1], [2], [40]–[54]). The CNN-based methods depend on the large-scale data set training to learn a nonlinear functional mapping between the low spatial resolution MS images and the high spatial resolution MS images. After the training phase, it is easy to predict/compute the pansharpened image by the learned nonlinear mapping. Masi *et al.* [41] first proposed a simple and effective CNN architecture with three layers called pansharpening neural network (PNN). This architecture is mainly based on a previous CNN architecture for single image super-resolution [55] and yields state-of-the-art pansharpening outcomes. Liu *et al.* [52] presented a good way to inject the high-pass details of the PAN image into the upsampled MS image, even by exploiting classical injection gains. This way is a bit like the scheme of traditional CS and MRA methods, but the extraction of the high-pass details is not in agreement with the classical procedures performed by CS and MRA approaches. Yang *et al.* [1] proposed a deeper network architecture than PNN, which is called PanNet. The PanNet architecture incorporates domain-specific knowledge and mainly focuses on two important issues, i.e., spectral and spatial preservations, obtaining state-of-the-art results. Furthermore, due to the use of high-pass filtering, the given architecture also shows the relevant ability of network generalization. He *et al.* [2] proposed a detail injection-based convolutional neural network (DiCNN). In particular, the authors developed two detail injection-based architectures, i.e., DiCNN1, whose detail injection depends on both MS and PAN images, and DiCNN2, whose detailed injection depends only on PAN images. DiCNN2 is designed to alleviate the computational burden; instead, DiCNN1 is more oriented to high quantitative performance getting state-of-the-art results. However, there is still room for improvement focusing on aspects as network complexity, training time, robustness, and so forth.

In this article, we propose deep CNNs to address the pansharpening problem, even accounting for fusion schemes proposed in the literature. In particular, we focus our attention on traditional CS and MRA frameworks. The details are extracted using these two philosophies. Instead, the nonlinear injection model is estimated through CNNs. These approaches are here named CS-Net and MRA-Net, respectively. Inspired by these solutions, we further investigate this idea feeding the network with details directly extracted by differencing the single PAN image with each MS band. This solution allows us to avoid compromising the spatial information with a preprocessing step using detailed extraction techniques proposed in classical pansharpening approaches, letting the CNN spectrally adjust the extracted details (e.g., the details are clearly biased) through the estimation of the nonlinear and local injection model. This approach will be called Fusion-Net from hereon. The proposed approaches are tested on several data sets acquired by the WorldView-2, WorldView-3, GaoFen (GF)-2, and QuickBird (QB) data sets. The experimental analysis is conducted both at reduced and full resolutions. The benchmark consists of state-of-the-art CS and MRA approaches and ML methods for pansharpening. The proposed Fusion-Net method clearly shows state-of-the-art performance outperforming the methods in the adopted benchmark both quantitatively and qualitatively. Finally, discussions about network complexity, training time, convergence, and robustness are provided to the readers for all the compared CNN approaches.

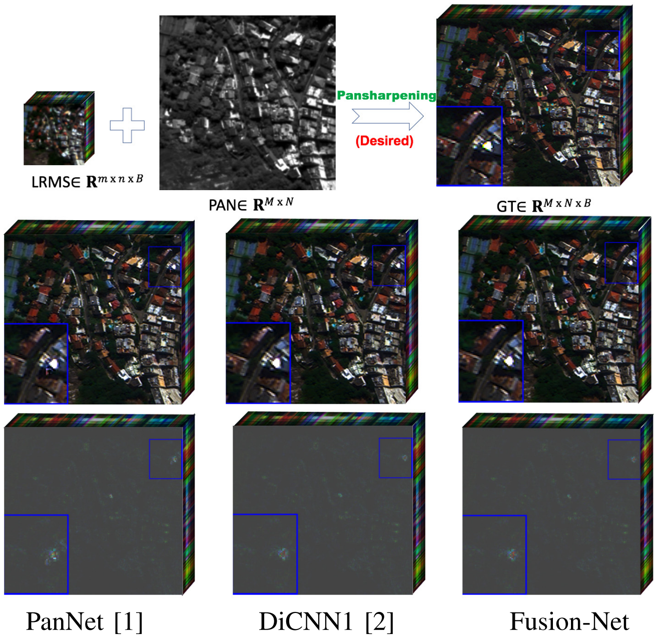

LRMS∈ $\mathbf{R}^{m\times n\times B}$
PAN∈ $\mathbf{R}^{M\times N}$
Pansharpening
(Desired)
GT∈ $\mathbf{R}^{M\times N\times B}$
PanNet [1]
DiCNN1 [2]
Fusion-Net

Fig. 1. First row: flowchart for pansharpening on an eight-band WorldView-3 satellite data with a spatial resolution factor equal to 4. The figure includes the low spatial resolution MS image, the PAN image, and the GT image. Second and third rows: the pansharpened images and the corresponding AEMs of three (high-performance) deep CNNs, i.e., PanNet (SAM/ERGAS/Q8 = 5.05/3.33/0.936) [1], DiCNN1 (5.02/3.22/0.945) [2], and the proposed Fusion-Net (4.63/3.02/0.951). In the second row, the fused products are represented in natural colors. From the third row, it is clear that the proposed Fusion-Net yields the darker AEM implying superior performance with respect to the competitors.

In summary, the main contributions of this work are as follows.

1) Two physically justified CNNs (i.e., CS-Net and MRA-Net) have been proposed deriving them from the traditional CS and MRA frameworks.
2) Inspired by CS-Net and MRA-Net, the Fusion-Net has also been proposed to reach state-of-the-art performance (see the comparison among the high-performance CNNs in Fig. 1 for a WorldView-3 data set). Moreover, the Fusion-Net has a simple architecture with fewer network parameters, thus resulting in more effective than some previously developed network architectures for pansharpening.

3) A broad experimental analysis has been provided based on several data sets. The performance is assessed both at reduced and full resolutions. The numerical outcomes are also corroborated by a qualitative analysis. Finally, a deep discussion on the network generalization, convergence property, computational time, and robustness on large data sets has been provided to the readers for all the considered CNN approaches.

This article is organized as follows. The related works and motivations are introduced in Section II. The proposed three network architectures will be detailed in Section III. Section IV is devoted to the description of the experimental results and the related discussions. Finally, conclusions are drawn in Section V.

II. Related Works and Motivations

The proposed network is initially inspired by two traditional pansharpening frameworks, i.e., CS and MRA. Therefore, we will first introduce them in this section, and then, we will move toward the motivations under the choice of the proposed network architectures.

A. CS

The general fusion equation for CS-based methods is as follows:

$$\widehat{\mathbf{MS}}_i = \widetilde{\mathbf{MS}}_i + g_i(\mathbf{P} - \mathbf{I}_L), \quad i = 1, 2, \ldots, B \tag{1}$$

where $\widehat{\mathbf{MS}}_i \in \mathbb{R}^{M \times N}$ is the ith band of the high spatial resolution MS image, $\widetilde{\mathbf{MS}}_i \in \mathbb{R}^{M \times N}$ is the ith band of the upsampled version of the low spatial resolution MS image, $g_i \in \mathbb{R}$ is the ith injection coefficient (a real number for global approaches) that controls the injection of the extracted details, $\mathbf{P} \in \mathbb{R}^{M \times N}$ represents the PAN image, and $\mathbf{I}_L \in \mathbb{R}^{M \times N}$ is the intensity component, generally defined as $\mathbf{I}_L = \sum_{i=1}^{B} \omega_i \widetilde{\mathbf{MS}}_i$, where $\omega_i \in \mathbb{R}$ is the ith weight. Many CS-based pansharpening algorithms rely upon (1), just changing the ways to estimate the injection coefficients g_i and the weights ω_i (see [9], [11]–[13]).

Equation (1) could be further rewritten in the following multiband form:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + \mathbf{g} \odot \left(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}}\right) \tag{2}$$

where $\widehat{\mathbf{MS}} \in \mathbb{R}^{M \times N \times B}$ and $\widetilde{\mathbf{MS}} \in \mathbb{R}^{M \times N \times B}$ are obtained by stacking the bands $\widehat{\mathbf{MS}}_i$, $i = 1, 2, \ldots, B$ and $\widetilde{\mathbf{MS}}_i$, $i = 1, 2, \ldots, B$, respectively, $\mathbf{P}^{\mathbf{D}} \in \mathbb{R}^{M \times N \times B}$ and $\mathbf{I}_L^{\mathbf{D}} \in \mathbb{R}^{M \times N \times B}$ are yielded by duplicating along the spectral dimension the PAN image, $\mathbf{P}$, and the intensity component, $\mathbf{I}_L$, respectively, $\mathbf{g} = (g_1, g_2, \ldots, g_B)^T \in \mathbb{R}^B$ is a vector of coefficients g_i as in (1), and $\odot$ is an operator indicating that the ith element of $\mathbf{g}$ multiplies the ith spectral band of $\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}}$.

B. MRA

Similar to the CS-based method, the MRA-based method follows the following equation:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + \mathbf{g} \odot \left(\mathbf{P}^{\mathbf{D}} - \mathbf{P}_L^{\mathbf{D}}\right) \tag{3}$$

where $\widehat{\mathbf{MS}}$, $\widetilde{\mathbf{MS}}$, $\mathbf{P}^{\mathbf{D}}$, $\mathbf{g}$, and the operator $\odot$ have the same definitions as in (2). Different from $\mathbf{I}_L^{\mathbf{D}}$ in (2), $\mathbf{P}_L^{\mathbf{D}} \in \mathbb{R}^{M \times N \times B}$ is yielded by duplicating along the spectral dimension of the $\mathbf{P}_L$ image that represents the low-pass spatial resolution version of the PAN image, $\mathbf{P}$. By differencing $\mathbf{P}^{\mathbf{D}}$ and $\mathbf{P}_L^{\mathbf{D}}$, i.e., $\mathbf{P}^{\mathbf{D}} - \mathbf{P}_L^{\mathbf{D}}$, the PAN spatial details can be extracted. Classical MRA approaches differ from each other in the way to extract PAN details and how to estimate the injection coefficient $\mathbf{g}$ in (3) (see [14], [15], [17]).

C. Motivations

The CS and MRA approaches have achieved promising performance in the field of pansharpening. However, a big limitation for both the classes is the common assumption of using linear injection models, which does not generally hold having a look at the relative spectral responses of sensors usually exploited for pansharpening (e.g., it is easy to note the overlaps among the MS spectral responses).

This consideration has motivated us to avoid linear injection models developing nonlinear approaches, aiming to replace the detailed injection phases in both CS and MRA methods. Deep convolutional neural networks (DCNNs) can easily manage this nonlinear mapping task due to the fact that they are able to reproduce strong nonlinearities in the data. Thus, they represent the best solution for the problem at hand. In particular, we still follow the general classical framework based on two phases: 1) detail extraction and 2) detail injection into the original MS image. However, we address the issue of nonlinear and local estimation of injection coefficients leveraging on DCNNs. Thus, in what follows, we will present the three proposed solutions based on different DCNN architectures for pansharpening (i.e., CS-Net, MRA-Net, and Fusion-Net).

III. Proposed Network Architectures

This section is devoted to the presentation of the DCNNs proposed in this work. We will present first the two CS- and MRA-based networks. Afterward, the Fusion-Net will be detailed.

A. CS-Net

Let us recall (2), in which the pansharpened product $\widehat{\mathbf{MS}}$ is equal to the sum of the upsampled MS image $\widetilde{\mathbf{MS}}$ and the injected details $\mathbf{g} \odot (\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}})$. In this equation, the upsampled MS image $\widetilde{\mathbf{MS}}$ holds the spatial information at low resolution, and $(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}})$ provides the high frequency details, injected through $\mathbf{g}$.

Equation (2) requires the estimation of the injection coefficients $\mathbf{g}$. Instead, we ignore that the injection coefficients considering the pansharpened image, $\widehat{\mathbf{MS}}$, consist of the upsampled MS image $\widetilde{\mathbf{MS}}$ plus the details coming from the nonlinear mapping provided by a DCNN feeding it with $(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}})$. In summary, the CS-Net can be summarized as follows:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + f_{\mathbf{\Theta}_{\mathrm{CS}}}\left(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}}\right) \tag{4}$$

where $f_{\mathbf{\Theta}_{\mathrm{CS}}}$ is the nonlinear mapping with the network parameter $\mathbf{\Theta}_{\mathrm{CS}}$ that could be learned from a large-scale training

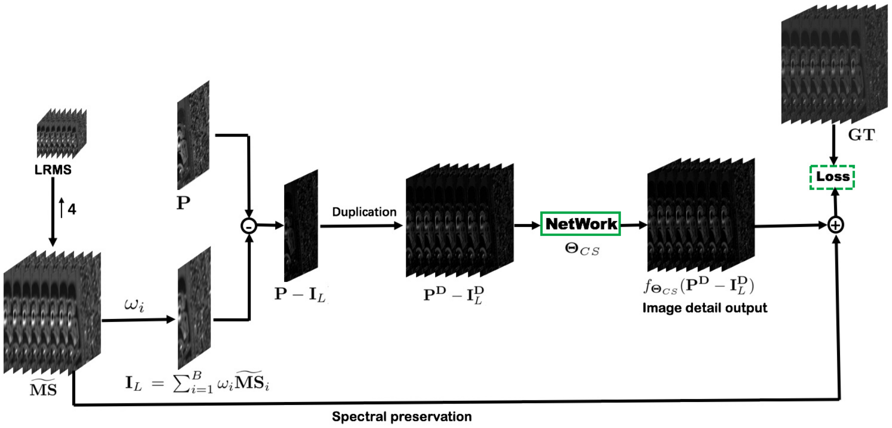

LRMS
↑4
P
Duplication
NetWork
$\mathbf{\Theta}_{CS}$
GT
Loss
ω_i
$\mathbf{P}-\mathbf{I}_L$
$\mathbf{P}^{\mathbf{D}}-\mathbf{I}_L^{\mathbf{D}}$
$f_{\mathbf{\Theta}_{CS}}(\mathbf{P}^{\mathbf{D}}-\mathbf{I}_L^{\mathbf{D}})$
Image detail output
$\widetilde{\mathbf{MS}}$
$\mathbf{I}_L = \sum_{i=1}^{B}\omega_i\widetilde{\mathbf{MS}}_i$
Spectral preservation

Fig. 2. Architecture of the CS-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork," please refer to Section III-D.

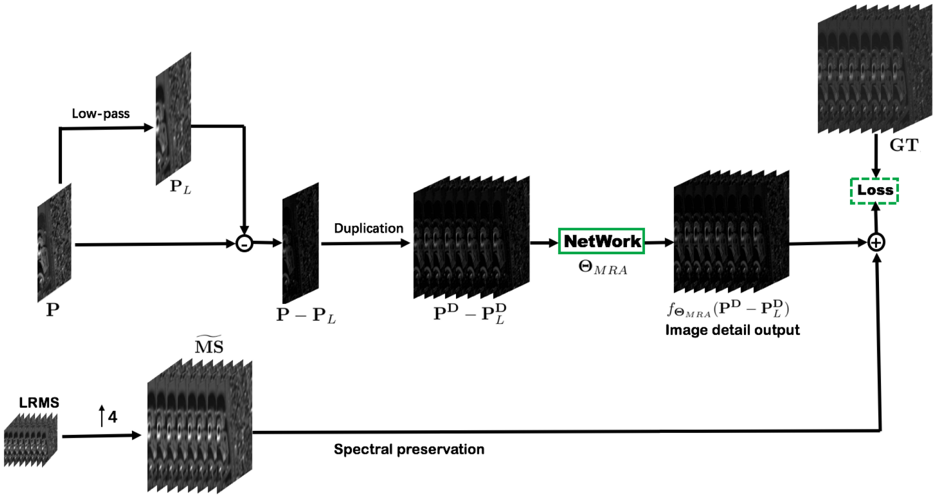

Low-pass
$\mathbf{P}_L$
P
Duplication
NetWork
$\mathbf{\Theta}_{MRA}$
GT
Loss
$\mathbf{P}-\mathbf{P}_L$
$\mathbf{P}^{\mathbf{D}}-\mathbf{P}_L^{\mathbf{D}}$
$f_{\mathbf{\Theta}_{MRA}}(\mathbf{P}^{\mathbf{D}}-\mathbf{P}_L^{\mathbf{D}})$
Image detail output
$\widetilde{\mathbf{MS}}$
LRMS
↑4
Spectral preservation

Fig. 3. Architecture of the MRA-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork," please refer to Section III-D.

data set. Several solutions to get the weights $\omega_i, i = 1, \ldots, B$ for $\mathbf{I}_L$ are given by the pansharpening literature, from constant to band-dependent (estimated) weights. From our broad experimental analysis, comparable results can be obtained by the CS-Net using these different intensity components.

Starting from (4), it is easy to build the corresponding network architecture (see Fig. 2). In particular, the final loss function for the CS-Net can be defined under the metric of mean squared error (MSE) computed on training examples. Hence, we have

$$\text{Loss}(\mathbf{\Theta}_{\text{CS}}) = \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{\mathbf{MS}}_{\{k\}} + f_{\mathbf{\Theta}_{\text{CS}}}\left(\mathbf{P}^{\mathbf{D}}_{\{k\}} - \mathbf{I}^{\mathbf{D}}_{L\{k\}}\right) - \mathbf{GT}_{\{k\}}\right\|_F^2 \tag{5}$$

where n represents the number of training examples, $\|\cdot\|_F$ is the Frobenius norm, and $\mathbf{GT}_{\{k\}}$ is the kth example extracted from the ground-truth (GT) image. By minimizing the loss function (5), the network $f_{\mathbf{\Theta}}$ will be enforced to automatically learn an optimal mapping with parameters $\mathbf{\Theta}_{\text{CS}}$. Thus, the fusion can be completed by summing the weighted spatial details to the upsampled MS image following (4).

B. MRA-Net

Similar to the analysis of the CS-Net, we derive the architecture of MRA-Net. The pansharpened image $\widehat{\mathbf{MS}}$ in (3) consists of the upsampled MS image $\widetilde{\mathbf{MS}}$ plus the injected details $\mathbf{g} \odot (\mathbf{P}^{\mathbf{D}} - \mathbf{P}^{\mathbf{D}}_L)$. Again, we ignore the injection coefficients $\mathbf{g}$ by imposing a nonlinear mapping function estimated through a DCNN fed by $(\mathbf{P}^{\mathbf{D}} - \mathbf{P}^{\mathbf{D}}_L)$. Therefore, the MRA-Net can be summarized as follows:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + f_{\mathbf{\Theta}_{\text{MRA}}}\left(\mathbf{P}^{\mathbf{D}} - \mathbf{P}^{\mathbf{D}}_L\right) \tag{6}$$

where $f_{\mathbf{\Theta}_{\text{MRA}}}$ is the nonlinear mapping function with network parameters $\mathbf{\Theta}_{\text{MRA}}$. Several solutions to get $\mathbf{P}_L$ from $\mathbf{P}$ are given by the pansharpening literature, from average to Gaussian filters. Again, from our broad experimental analysis, comparable results can be obtained by the MRA-Net using these different ways to spatially filter the PAN image.

Using (6), it is easy to design the network architecture of the MRA-Net (see Fig. 3). In particular, the loss function of

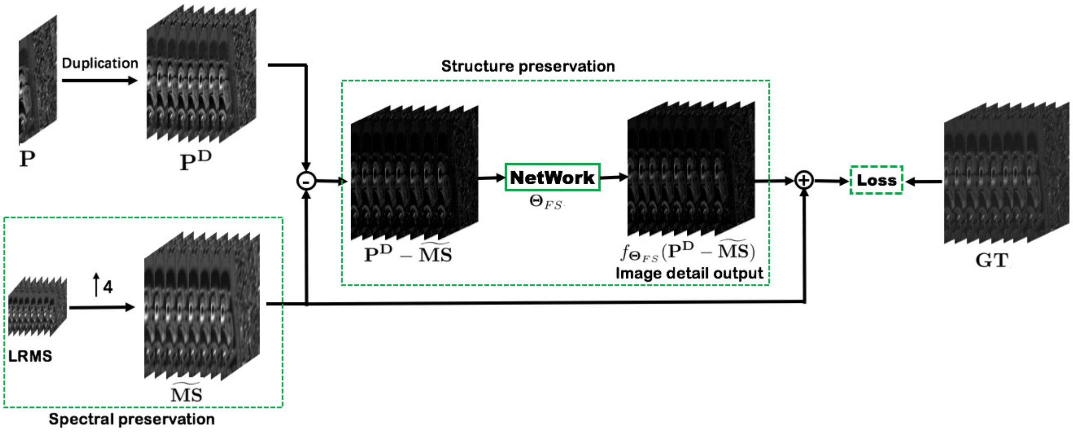

Duplication
P
$\mathbf{P}^{\mathrm{D}}$
Structure preservation
NetWork
Θ_{FS}
$\mathbf{P}^{\mathrm{D}} - \widetilde{\mathbf{MS}}$
$f_{\Theta_{FS}}(\mathbf{P}^{\mathrm{D}} - \widetilde{\mathbf{MS}})$
Image detail output
Loss
GT
↑4
LRMS
$\widetilde{\mathbf{MS}}$
Spectral preservation

Fig. 4. Architecture of the Fusion-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork," please refer to Section III-D.

the MRA-Net is defined as follows:

$$\mathrm{Loss}(\boldsymbol{\Theta}_{\mathrm{MRA}}) = \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{\mathbf{MS}}_{\{k\}} + f_{\boldsymbol{\Theta}_{\mathrm{MRA}}}\left(\mathbf{P}^{\mathbf{D}}_{\{k\}} - \mathbf{P}^{\mathbf{D}}_{L\{k\}}\right) - \mathbf{GT}_{\{k\}}\right\|_F^2 \tag{7}$$

where the definitions of symbols are the same as that of (5).

C. Fusion-Net

CS-Net and MRA-Net have been developed starting from the two classical fusion schemes related to CS and MRA. Thus, they have a solid and physical justification rooted in the pansharpening literature. However, in order to extract the details, preliminary assumptions should be done either on the shape of the spatial filters (for MRA-Net) or on the spectral model ruling the projection of the MS image into the PAN domain (for CS-Net). Errors in this phase can have a great impact on the outcomes reducing the performance of the proposed approaches. Thus, aiming of having a detail-based architecture, but avoiding the above-mentioned issue, the solution of subtracting the duplicated version of the PAN image, $\mathbf{P}^{\mathbf{D}}$, with the upsampled MS image, $\widetilde{\mathbf{MS}}$, is advisable. This has also the advantage to alleviate the computation burden of the approach avoiding to calculate $\mathbf{I}^{\mathbf{D}}_L$ or $\mathbf{P}^{\mathbf{D}}_L$. The limitation of this solution is instead related to the strong spectral distortion introduced in the extracted details (e.g., biased details) that can be easily compensated by the network during its training phase.

Another clear issue in the design of CS-Net and MRA-Net is that only data projected into the PAN domain are presented to the DCNNs. Namely, the inputs of the networks are practically monochromatic images i.e., without any spectral content). Thus, both the CS-Net and the MRA-Net receive no spectral information from these data. The networks fed in this way are not able to adequately reconstruct image features along the spectral direction, even training them with enough examples and a proper number of iterations. Instead, the use of $\mathbf{P}^{\mathbf{D}} - \widetilde{\mathbf{MS}}$ as details to feed the network has the advantage to intrinsically introduce the spectral information. All these cues are supported by the experimental analysis showing that the Fusion-Net outperforms the other two proposed approaches.

Similar to the CS-Net and the MRA-Net, we ignore the injection coefficients $\mathbf{g}$ in the general fusion equation of CS/MRA methods, allowing a DCNN to automatically estimate the nonlinear injection model. The Fusion-Net can be summarized as follows:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + f_{\boldsymbol{\Theta}_{\mathrm{FS}}}\left(\mathbf{P}^{\mathbf{D}} - \widetilde{\mathbf{MS}}\right) \tag{8}$$

where $f_{\boldsymbol{\Theta}_{\mathrm{FS}}}$ is the nonlinear mapping with network parameters $\boldsymbol{\Theta}_{\mathrm{FS}}$.

Starting from (8), the network architecture of the proposed Fusion-Net is described in Fig. 4. In particular, the loss function for the Fusion-Net is as follows:

$$\mathrm{Loss}(\boldsymbol{\Theta}_{\mathrm{FS}}) = \frac{1}{n}\sum_{k=1}^{n}\left\|\widetilde{\mathbf{MS}}_{\{k\}} + f_{\boldsymbol{\Theta}_{\mathrm{FS}}}\left(\mathbf{P}^{\mathbf{D}}_{\{k\}} - \widetilde{\mathbf{MS}}_{\{k\}}\right) - \mathbf{GT}_{\{k\}}\right\|_F^2 \tag{9}$$

where the definitions of symbols are the same as that of (5).

Note that the Fusion-Net proposed in the work can be also regarded as a support strategy for deep learning (DP), thus improving the performance of existing methods. Please refer to Section IV for details about the performance gains.

D. Network Selection

We have proposed three deep network architectures for pansharpening, i.e., CS-Net, MRA-Net, and Fusion-Net. They all involved a subnetwork for training, i.e., "NetWork" (see the solid green boxes in Figs. 2–4). The main structure of "NetWork" is presented in Fig. 5(a). Wherein, we choose an effective network recently proposed in the literature, called ResNet [56], as the subnetwork of the proposed architectures since the ResNet can bring the conventional CNN to deeper layers leading to an effective and competitive performance in many image applications. Fig. 5 shows the basic structure of one ResNet block, in which one skip connection for every two convolutional layers is shown. In practical experiments, we need to empirically tune the number of ResNet blocks to control the final convolutional layers, aiming to achieve the best performance (see the parameter setting in Section IV).

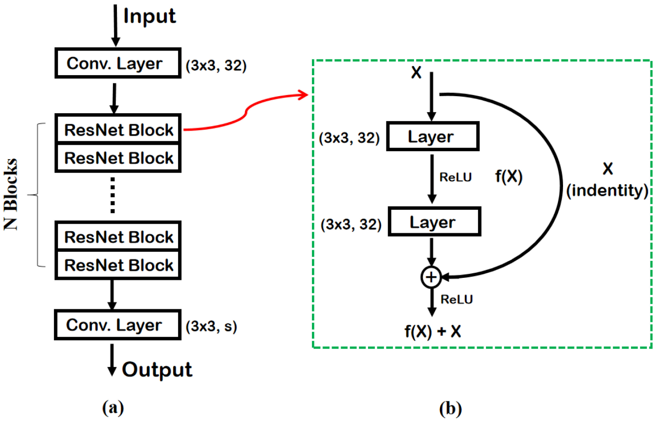


Fig. 5. (a) Structure of "NetWork" with several ResNet blocks (see the solid green boxes in Figs. 2–4). Note that $(3 \times 3, 32)$ represents 32 convolution kernels with size 3×3, and s depends on the number of MS bands (e.g., for four-band image $s = 4$ and for eight-band image $s = 8$). (b) Details of one ResNet block [56] that is used in our architectures. Each ResNet block contains two nonlinear rectified linear unit (ReLU) activation functions. In particular, the ResNet block is slightly different in the case of "MRA-Net," where we have $(3 \times 3, 64)$. For further details, please refer to Table I.

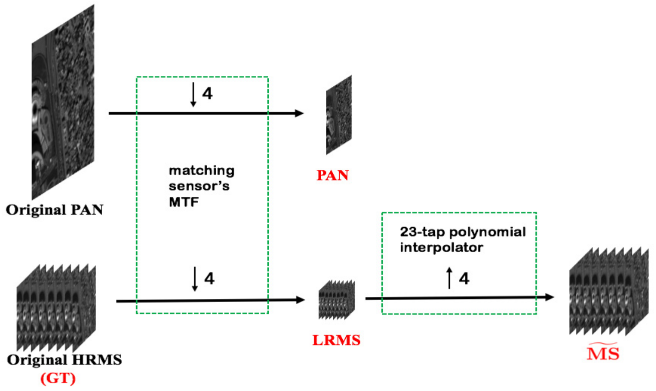


Fig. 6. Generation process of the training data set by Wald's protocol. Note that the data indicated with the red text are the generated training data used to feed the networks, i.e., the GT, the low spatial resolution MS (LRMS) image, the PAN, and the upsampled MS image ($\widetilde{\mathbf{MS}}$).

E. Generation of Training Data

In this work, we train the CNNs on WorldView-3 (eight bands) satellite data sets that can be easily downloaded on the public website.[1] After downloading the data sets, we simulate 12 580 PAN/MS/GT image pairs with the size 64×64, $16 \times 16 \times 8$, and $64 \times 64 \times 8$, respectively, and then split them into 70/20/10% for training (8806 examples[2])/validation (2516 examples)/testing (1258 examples). Note that since the GT images are not available, we need to follow Wald's protocol [57] to get them. The process of simulating the training data set by Wald's protocol is illustrated in Fig. 6. It mainly contains the following steps: 1) downsampling the original PAN and the original MS image by a resolution factor 4 using modulation transfer function (MTF)-based filters, seeing the downsampled PAN image as the training PAN image and the downsampled MS image as the training MS image; 2) taking the original MS image as the training GT image; and 3) upsampling the training MS image by using a polynomial kernel with 23 coefficients [17] and interpreting the output as the upsampled MS image. Following steps 1–3, it is easy to generate the training data. The validation and testing data sets are similarly built.

IV. Experimental Results

In this section, we compare the proposed network architectures (i.e., CS-Net, MRA-Net, and Fusion-Net) with some recent state-of-the-art pansharpening approaches belonging to the CS, the MRA, and the ML classes. First, the employed sensors, the benchmark, and the adopted quality indexes will be described. Afterward, the experimental analysis both at reduced and full resolutions will be described.

A. Data Sets

Several data sets have been acquired by the WorldView-2 and WorldView-3 sensors. The former provides a high-resolution PAN channel and eight MS bands. Four standard colors (red, green, blue, and near-infrared 1) and four new bands (coastal, yellow, red edge, and near-infrared 2) are acquired. Although the native spatial resolution would be greater, the images are distributed with a pixel size of 0.5 and 2 m for PAN and MS, respectively. The spatial resolution ratio is equal to 4. The radiometric resolution is 11 bits. WorldView-3 data have the same features as WorldView-2 data, but with a spatial resolution of about 0.3 m for the PAN channel and of about 1.2 m for the MS bands and a radiometric resolution of 11 bits. Moreover, we also assess the performance on four-band (red, green, blue, and near-infrared) data sets. In particular, QB data are considered having a spatial resolution of 2.4 and 0.61 m for the MS and PAN images, respectively, and a radiometric resolution of 11 bits. Finally, images acquired by the GF-2 sensor have been exploited with a spatial resolution of 3.2 and 0.8 m for the MS and PAN images, respectively, and a radiometric resolution of ten bits (please see Section IV-H for more details).

B. Benchmark

The proposed benchmark consists of the following methods: the MS image interpolation using a polynomial kernel with 23 coefficients (EXP) [17], the GS sharpening approach [12], the SFIM [13], the PRACS approach [11], the BDSD method [9], the robust BDSD approach (BDSD-PC) [10], the GLP with MTF-matched filter [58] and multiplicative injection model [59] [GLP-high-pass modulation (HPM)], the GLP with MTF-matched filter [58] and regression-based injection model [GLP-context-based decision (CBD)] [3], [17], the GLP with full-scale regression (GLP-Reg) [20], the state-of-the-art CNN-based method for pansharpening (PNN) [41],[3]

[1] http://www.digitalglobe.com/samples?search=Imagery

[2] We tried to simulate the same training data set as in [1] (PanNet), but, in the original article, the authors do not indicate which WorldView-3 data sets are selected for the training. However, in our work, all deep learning-based methods are trained on the same data set for a fair comparison.

[3] Note that the given source code in open remote sensing does not contain the trained models for WV2 and WV3; thus, we reimplemented the network with default parameters in Python using Tensorflow for simplicity of comparison.

TABLE I

OPTIMAL PARAMETERS FOR THE COMPARED DCNNS. NOTATION: **ITER. #** (ITERATION NUMBER), **BS** (MINIBATCH SIZE), **ALGO** (OPTIMIZATION ALGORITHM), **LR** (LEARNING RATE), **FS** (FILTER SIZE FOR EACH LAYER), **FILT. #** (FILTER NUMBER FOR EACH LAYER), **N** (THE NUBMER OF RESNET BLOCKS), AND **LY. #** (NUMBER OF LAYERS)

Para.	**PNN**	**DRPNN**	**CS-Net**	**MRA-Net**	**DiCNN1**	**PanNet**	**DMDNet**	**Proposed**
Iter. #	1.12×10^6	3×10^5	1.8×10^5	1.6×10^5	3×10^5	2.4×10^5	2.5×10^5	1.4×10^5
Bs	128	64	64	32	64	32	32	32
Algo	SGD	SGD	Adam	Adam	Adam	Adam	Adam	Adam
Lr	0.00001	0.05, 0.005	0.0003	0.0003	0.0001	0.0001	0.0001	0.0003
Fs	9×9, 5×5	7×7	3×3	3×3	3×3	3×3	3×3	3×3
Filt. #	64, 32	64	32	64	64	32	64	32
N	-	-	4	8	-	4	4	4
Ly. #	3	11	10	18	3	10	10	10

the state-of-the-art CNN-based method for pansharpening deep residual PNN (DRPNN) [44],[4] the state-of-the-art CNN-based method for pansharpening (PanNet) [1],[5] the state-of-the-art CNN-based method for pansharpening (DiCNN1) [2],[6] the state-of-the-art CNN-based method with dilated convolution for pansharpening (DMDNet) [60],[7] and the proposed CS-Net, MRA-Net, and Fusion-Net. Note that the source codes of all CS and MRA-based methods can be found on public websites.[8]

For a fair comparison, all the compared CNNs are trained on Python 3.5.2 with Tensorflow 1.0.1 on a desktop PC equipped with a GPU NVIDIA GeForce GTX 1080 with 8 GB.

C. Quality Assessment

The performance assessment is conducted both at reduced and full resolutions. The former is performed using the spectral angle mapper (SAM) [61], the relative dimensionless global error in synthesis (ERGAS) [62], the spatial correlation coefficient (SCC) [63], and the universal image quality index for four-band images (Q4) and eight-band images (Q8) [64]. In particular, the ideal value for Q4, Q8, and SCC is 1, while, for SAM and ERGAS, it is 0. Furthermore, to evaluate the performance at full resolution, we employ the quality without reference (QNR), D_λ, and D_s indexes [6]. The QNR has an ideal value of 1, instead D_λ and D_s have an ideal value of 0.

D. Parameters Tuning

Before going through the description of the experimental results, the tuning parameters of the CNN-based approaches are shown. As mentioned in Section III-E, the training data for PanNet and DiCNN1 in this work are different from that of their original articles; thus, it may lead to slightly different optimal parameters. We tried to do our best to have the highest performance for both the PanNet and the DiCNN1 with a full parameter tuning in order to have a fair comparison. We summarize the optimal parameters of all CNN methods in Table I.[9]

TABLE II

QUANTITATIVE COMPARISON OF THE COMPARED DEEP NETWORKS FOR THE TESTING DATA SET THAT INCLUDES 1258 SAMPLES. BEST RESULTS IN BOLDFACE

	SAM (± std)	**ERGAS** (± std)	**Q8** (± std)	**SCC** (± std)
PNN	4.4015 ± 1.3292	3.2283 ± 1.0042	0.8883 ± 0.1122	0.9215 ± 0.0464
DRPNN	4.2657 ± 1.2431	3.0317 ± 0.9507	0.9010 ± 0.1089	0.9317 ± 0.0475
DiCNN1	3.9805 ± 1.3181	2.7367 ± 1.0156	0.9096 ± 0.1117	0.9517 ± 0.0471
PanNet	4.0921 ± 1.2733	2.9524 ± 0.9778	0.8941 ± 0.1170	0.9494 ± 0.0460
DMDNet	3.9714 ± 1.2482	2.8572 ± 0.9663	0.9000 ± 0.1141	0.9527 ± 0.0446
CS-Net	4.4851 ± 1.4605	3.1036 ± 1.1241	0.8937 ± 0.1156	0.9388 ± 0.0509
MRA-Net	4.5309 ± 1.4350	3.2657 ± 1.1169	0.8865 ± 0.1180	0.9372 ± 0.0482
Fusion-Net	**3.7435** ± **1.2259**	**2.5679** ± **0.9442**	**0.91353** ± **0.1122**	**0.9580** ± **0.0450**

E. Reduced Resolution Assessment

After the training phase, we need to validate the performance of the compared CNN methods on WorldView-3 testing data. In this phase, we exclude classical CS and MRA methods because they will be strongly penalized by the absence of a training phase using similar samples that will be found in the testing data set. Thus, this analysis is only devoted to comparing CNN-based approaches trained on the same examples.

In Table II, we first show the average quantitative results of the different methods on the testing data set containing 1258 testing examples. For each testing example, the sizes of PAN, MS, and GT images are the same as that of the training examples, i.e., 64 × 64 for the PAN image, 16 × 16 × 8 for the original low spatial resolution MS image, and 64 × 64 × 8 for the GT image. From Table II, it is clear that the proposed Fusion-Net obtains the best average quantitative performance for all the quality indexes. Furthermore, the standard deviations (std) of all the metrics get the smallest values for all the indexes, which also demonstrates the robustness of the proposed Fusion-Net. In particular, having a look at Table II,

[4] It is not easy to find the source code; thus, we reimplemented the network with default parameters in Python using Tensorflow for simplicity of comparison.

[5] Code link: https://xueyangfu.github.io/

[6] DiCNN1 has been implemented by ourselves.

[7] DMDNet has been implemented by ourselves.

[8] http://openremotesensing.net/kb/codes/pansharpening/

[9] Note that PanNet, CS-Net, MRA-Net, and Fusion-Net use ResNet blocks if the number of layers for one of these networks is 10, which means that there are 4 ResNet blocks (each block with two layers) and two extra input and output layers.

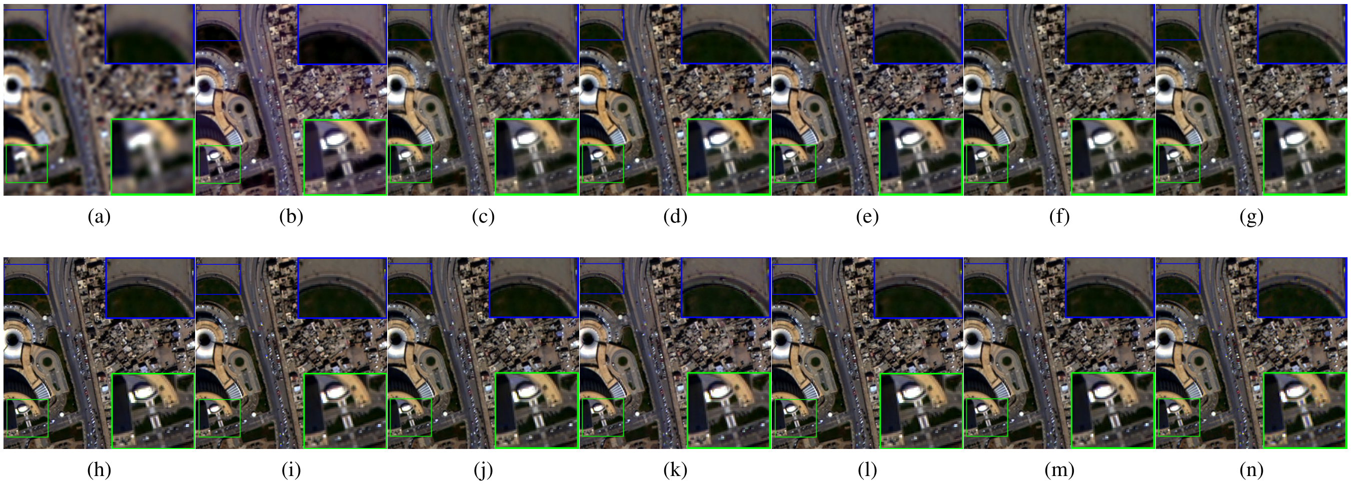


Fig. 7. Visual comparisons in natural colors of the most representative 13 approaches on the Rio data set (WorldView-3). (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

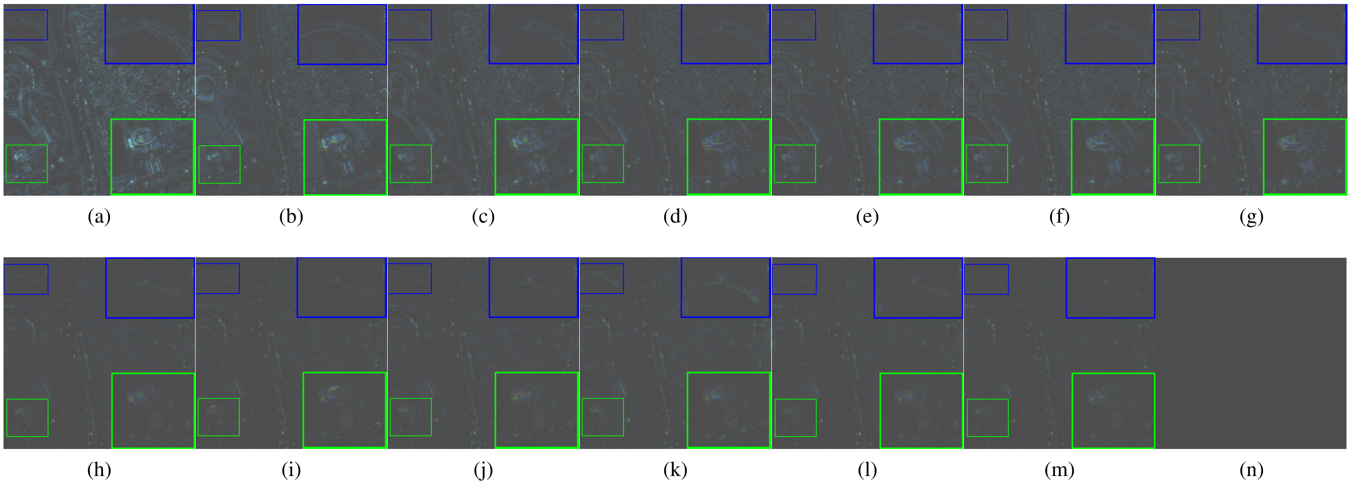


Fig. 8. AEMs of Fig. 7. (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

it is clear that the results of CS-Net and MRA-Net are worse than that of the recent DL-based methods (see the reasons underlined in the first two paragraphs in Section III-C). Hence, we will not show the results of CS-Net and MRA-Net from hereon, considering as a unique comparison the one with Fusion-Net. However, the presentation of CS-Net and MRA-Net is still meaningful since the proposed Fusion-Net is inspired and motivated by them.

A further test is about the use of two new WorldView-3 data sets capturing scenarios never presented to the networks in their training phase. In this case, the whole benchmark is used considering the comparison fair even when classical CS and MRA approaches are used. Again, Wald's protocol is used to generate a reference (GT) image, as described in Section III-E. The two data sets will be named Rio and Tripoli from hereon, which both hold 30-cm resolution. Their size is $256 \times 256 \times 8$ for the GT image, 256×256 for the PAN image, and $64 \times 64 \times 8$ for the original low spatial resolution MS image. Table III indicates that the best performance is still reached by the proposed Fusion-Net outperforming the performance of all the other compared pansharpening approaches for all the quality metrics. Similar conclusions can be drawn when the Tripoli data set is used.

The visual analysis further corroborates these numerical results. Indeed, in Fig. 7 (Rio data set), it is clear to see that the visual results provided by the classical CS and MRA methods (e.g., GS, SFIM, BDSD, BDSD-PC, GLP-Reg, and GLP-CBD) show low spatial performance with evident blur effects. Moreover, all six CNN methods perform significantly better than the classical methods (both spatially and spectrally). This demonstrates the ability of CNN methods to address the problem of pansharpening. It is worth to be remarked that it is not easy to distinguish the visual differences among the CNN methods in Fig. 7. This is due to the limitations in representing 8-bit RGB images instead of the 11-bit MS data. However, exploiting the calculation of the absolute error maps (AEMs) of Fig. 8, the visual advantages of the proposed Fusion-Net are pointed out getting lower

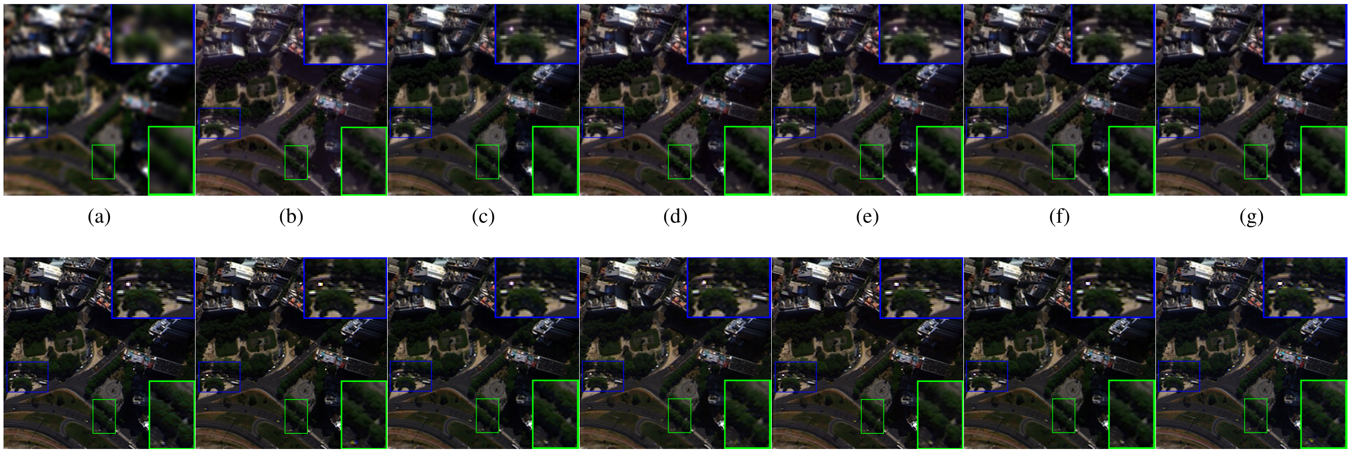


Fig. 9. Visual comparisons in natural colors of the most representative 13 approaches on the Tripoli data set (WorldView-3). (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

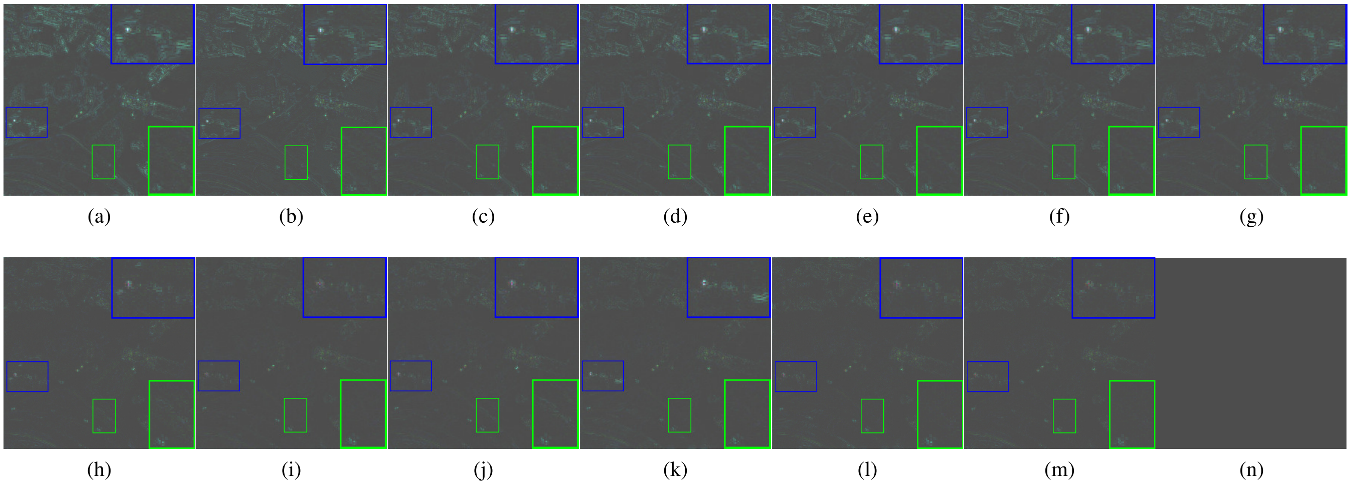


Fig. 10. AEMs of Fig. 9. (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

image residuals (see the close-up boxes in Fig. 8). The same conclusions can be drawn for the visual analysis of the fusion outcomes using the Tripoli data set in Figs. 9 and 10.

F. Full-Resolution Assessment

In this section, we test the performance of the proposed benchmark at the original (full) scale. In this case, the GT image is not available requiring quality indexes without reference for performance assessment purposes. We exploited 30 image pairs (MS and PAN) of WorldView-3 data at the original scale (OS) for testing the approaches using the QNR as the quality index. Table IV shows the quantitative assessment for all the methods in the benchmark. The six deep networks, i.e., PNN, DRPNN, DiCNN1, PanNet, DMDNet, and the proposed Fusion-Net, outperform the classical approaches. Having a look at the overall quality index QNR, the best average performance is obtained by the proposed Fusion-Net, even with a limited standard deviation implying that we got a robust result. The same can be stated for the spectral index D_λ. Moreover, the best performance (comparable with the PanNet one) is obtained on the spatial index D_s. Finally, Fig. 11 shows the visual performance on a full-resolution WorldView-3 data set, here named the Tripoli-OS data set. It is easy to remark from Fig. 11 the lower spatial performance of the classical CS and MRA methods (e.g., GS, SFIM, BDSD, BDSD-PC, GLP-Reg, and GLP-CBD), whereas all the CNN-based methods significantly outperform the classical approaches, both spatially and spectrally. Furthermore, the proposed Fusion-Net obtains better spatial performance than that of the other five CNN-based methods. Meanwhile, Fusion-Net is also able to preserve the spectral information.

G. Network Generalization

We have demonstrated that the proposed Fusion-Net outperforms the other pansharpening approaches in the benchmark on WorldView-3 data when the networks are also trained on WorldView-3 data. In this section, we will focus on the capability of the networks to generalize the results

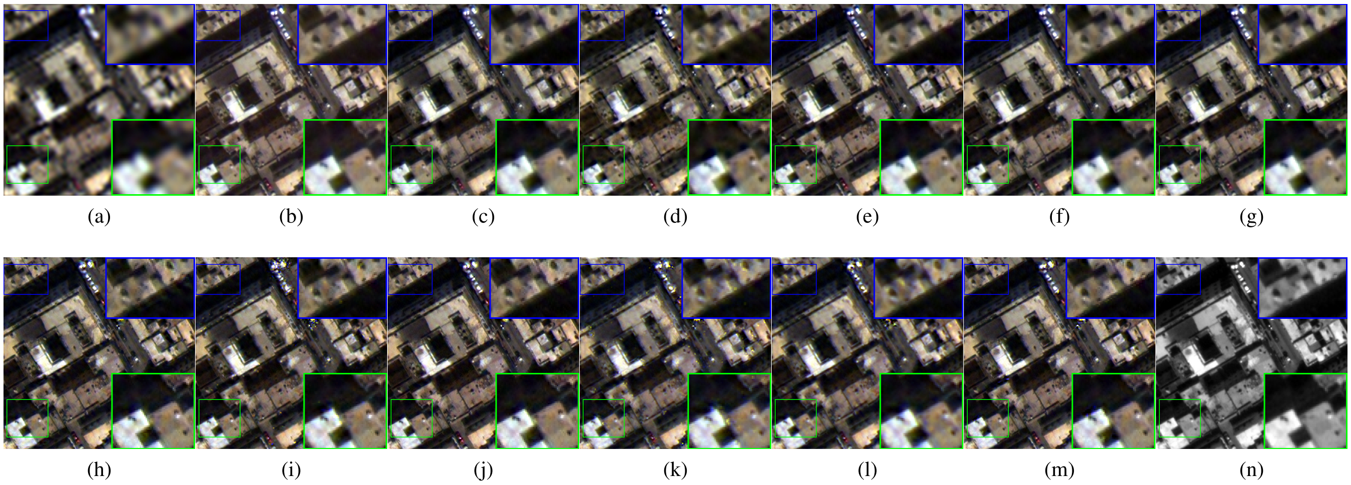


Fig. 11. Visual comparisons in natural colors of the most representative 13 approaches on the Tripoli-OS data set (WorldView-3) at the OS. (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) PAN.

TABLE III

QUANTITATIVE RESULTS FOR RIO DATA SET AND TRIPOLI DATA SET (WORLDVIEW-3). BEST RESULTS ARE IN BOLDFACE

	SAM	ERGAS	Q8	SCC	Time
Rio dataset					
EXP	4.203	5.5976	0.6927	0.6156	0.0312
GS	4.0614	3.8956	0.8666	0.8979	0.0440
SFIM	3.9132	3.563	0.8859	0.888	**0.0251**
BDSD	3.9567	2.8494	0.9361	0.9077	0.0796
BDSD-PC	3.8065	2.8494	0.9363	0.9061	0.1701
PRACS	4.026	3.2501	0.9062	0.8972	0.1765
GLP-HPM	4.1349	3.4917	0.8935	0.8817	0.2037
GLP-CBD	3.7068	2.7732	0.935	0.9092	0.1069
GLP-Reg	3.6871	2.776	0.9345	0.9095	0.1476
PNN	3.3728	2.3082	0.9488	0.9409	0.5475
DRPNN	3.1216	2.1669	0.9674	0.9585	0.6163
DiCNN1	3.0248	1.9119	0.9686	0.9627	0.5527
PanNet	3.0054	1.9506	0.9651	0.964	0.5880
DMDNet	2.9355	1.8119	0.96905	0.96993	0.6198
Fusion-Net	**2.8338**	**1.7510**	**0.9728**	**0.9714**	0.5477
Tripoli dataset					
EXP	6.7883	8.5719	0.7235	0.5129	0.0339
GS	7.1416	7.3237	0.7879	0.7251	0.0507
SFIM	6.3486	6.8407	0.8343	0.7341	**0.0231**
BDSD	6.8533	6.7863	0.8448	0.7338	0.0621
BDSD-PC	6.4985	6.7186	0.8475	0.7313	0.1615
PRACS	6.6680	7.0012	0.8266	0.7253	0.1848
GLP-HPM	6.8196	6.8881	0.8393	0.7350	0.1918
GLP-CBD	6.4178	6.5443	0.8503	0.7392	0.1102
GLP-Reg	6.4100	6.5463	0.8548	0.7394	0.1405
PNN	5.0778	3.9614	0.9214	0.9242	0.5515
DRPNN	4.8411	3.7810	0.9454	0.9468	0.6173
DiCNN1	4.7552	3.4978	0.9444	0.9482	0.5476
PanNet	4.6079	3.4227	0.9395	0.9516	0.5812
DMDNet	4.4282	3.1972	0.9458	0.9613	0.6020
Fusion-Net	**4.2764**	**3.0568**	**0.9522**	**0.9646**	0.5467

TABLE IV

AVERAGE VALUES OF QNR, D_λ, AND D_s WITH THE RELATED STANDARD DEVIATIONS (STD) FOR THE 30 FULL-RESOLUTION DATA (WORLDVIEW-3). BEST RESULTS ARE IN BOLDFACE

	QNR (± std)	D_λ (± std)	D_s (± std)
EXP	0.8032 ± 0.0612	0.0422 ± 0.0204	0.1241 ± 0.0661
GS	0.8866 ± 0.0606	0.0218 ± 0.0194	0.0944 ± 0.0458
SFIM	0.9234 ± 0.0523	0.0268 ± 0.0270	0.0518 ± 0.0292
BDSD	0.8822 ± 0.0286	0.0354 ± 0.0169	0.0852 ± 0.0264
BDSD-PC	0.8901 ± 0.0232	0.0344 ± 0.0152	0.0837 ± 0.0231
PRACS	0.8985 ± 0.0634	0.0224 ± 0.0194	0.0817 ± 0.0482
GLP-HPM	0.8834 ± 0.0323	0.0368 ± 0.0371	0.0718 ± 0.0492
GLP-CBD	0.9048 ± 0.0683	0.0333 ± 0.0285	0.0651 ± 0.0454
GLP-Reg	0.9082 ± 0.0601	0.0322 ± 0.0295	0.0629 ± 0.0521
PNN	0.9342 ± 0.0481	0.0297 ± 0.0232	0.0361 ± 0.0244
DRPNN	0.9437 ± 0.0630	0.0225 ± 0.029	0.0318 ± 0.0270
DiCNN1	0.9390 ± 0.0417	0.0214 ± 0.0210	0.0409 ± 0.0242
PanNet	0.9511 ± 0.0306	0.0221 ± 0.0137	0.0241 ± 0.0180
DMDNet	0.9587 ± 0.0310	0.0240 ± 0.0138	**0.0237 ± 0.0145**
Fusion-Net	**0.9612 ± 0.0272**	**0.0180 ± 0.0158**	0.0243 ± 0.0151

fusing data acquired by different sensors. To this aim, we exploit another data set acquired by another eight-band sensor, i.e., WorldView-2, but using the networks trained on WorldView-3 data. In order to have an accurate assessment, we still leverage on Wald's protocol to generate the so-called Stockholm data set acquired by the WorldView2 sensor. Quantitative results reported in Table V indicate that the proposed Fusion-Net is again the best approach outperforming the benchmark on the metrics of ERGAS and Q8. The DMDNet obtains slightly better SAM and SCC metrics than Fusion-Net since it employs the dilated convolution that could significantly increase the receptive field, whereas our Fusion-Net only uses conventional convolution. Fig. 12 corroborates this statement. It is easy to see that all the CNN methods yield better spatial performance than the CS and MRA approaches. In Fig. 13, the AEMs of Fig. 12 are also shown. Again, the proposed Fusion-Net exhibits the darker residual map demonstrating its superiority with respect to the other compared approaches even from a qualitative point of view.

H. Assessment on Four-Band Data Sets

In this section, we will extend the performance assessment to four-band data sets, i.e., acquired by the GF-2 and the QB sensors.

About the data simulation, we also follow the way described in Section III-E to generate the training and testing data.

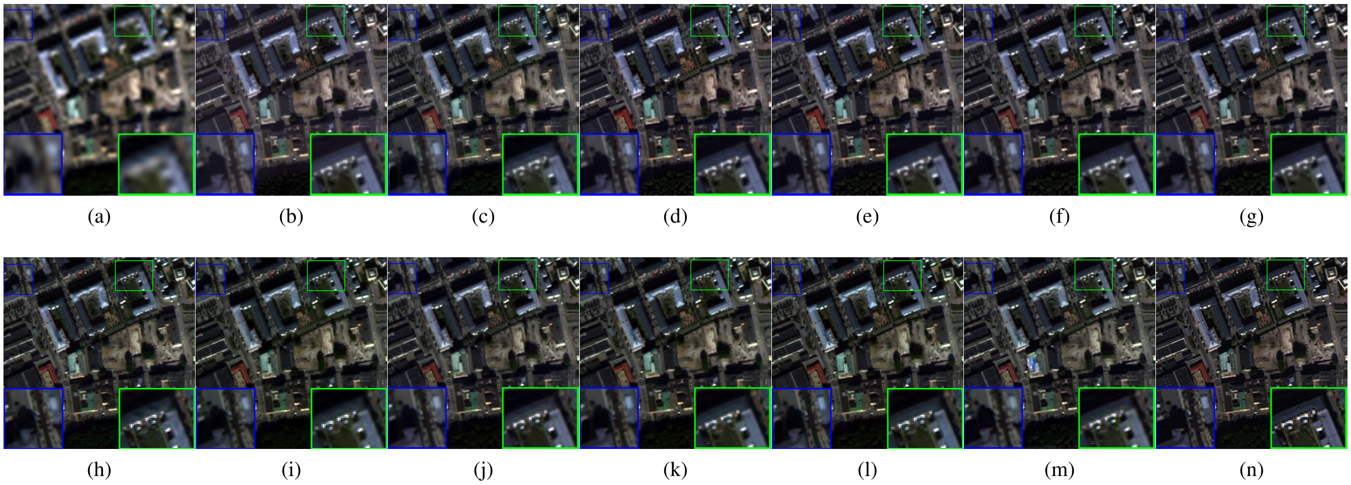


Fig. 12. Visual comparisons in natural colors of the most representative 13 approaches on the Stockholm data set (WorldView2). (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

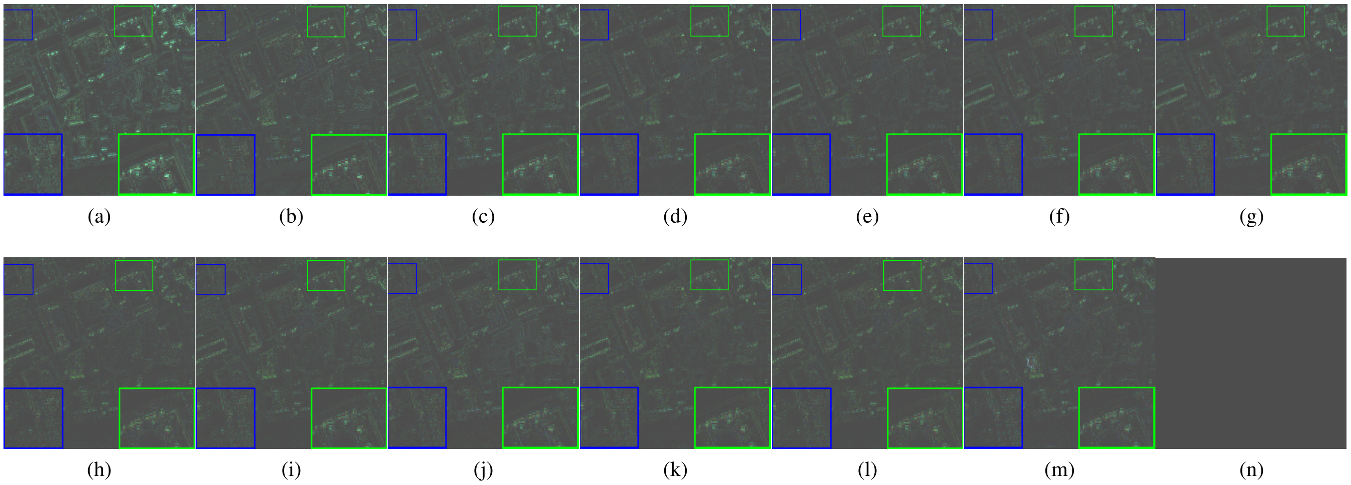


Fig. 13. AEMs of Fig. 12. (a) EXP. (b) GS. (c) SFIM. (d) BDSD. (e) BDSD-PC. (f) GLP-Reg. (g) GLP-CBD. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

TABLE V

QUANTITATIVE RESULTS ON THE STOCKHOLM DATA SET (WORLDVIEW2). BEST RESULTS ARE IN BOLDFACE

	SAM	**ERGAS**	**Q8**	**SCC**
EXP	7.8500	9.6793	0.6540	0.4505
GS	7.7296	7.3644	0.8075	0.8439
SFIM	7.1147	6.9570	0.8434	0.8562
BDSD	7.1824	6.3772	0.8798	0.860
BDSD-PC	7.0953	6.3233	0.8819	0.8578
PRACS	7.5894	7.4080	0.8314	0.8125
GLP-HPM	7.2988	6.9965	0.8527	0.8355
GLP-CBD	7.1098	6.5434	0.8752	0.8457
GLP-Reg	7.1195	6.4998	0.8776	0.8453
PNN	6.8624	5.6259	0.8642	0.8539
DRPNN	6.4798	5.6459	0.8843	0.8668
DiCNN1	6.8159	5.9773	0.8802	0.8797
PanNet	6.3916	5.6302	0.8897	0.8895
DMDNet	**6.1986**	5.5692	0.8903	**0.8965**
Fusion-Net	6.2784	**5.5499**	**0.8969**	0.8897

For the QB test case, we downloaded a large data set ($4906 \times 4906 \times 4$) acquired over the city of Indianapolis cutting it into two parts. The left part ($4906 \times 3906 \times 4$) is used to simulate 20 685 training samples (size: $64 \times 64 \times 4$), and the right part ($4906 \times 1000 \times 4$) is used to simulate 48 testing data (size: $256 \times 256 \times 4$). For the GF-2 test case, we downloaded a large data set ($6907 \times 7300 \times 4$) over the city of Beijing from the website[10] to simulate 21 607 training examples (size: $64 \times 64 \times 4$). Besides, a huge image acquired over the Guangzhou city is downloaded to simulate 81 testing data (size: $256 \times 256 \times 4$).

Figs. 14 and 15 present the visual performance of the five representative CNN-based methods.[11] The visual results provided by the six CNN methods all obtain competitive outcomes, both spatially and spectrally. As previously said, the RGB images shown in the first rows of Figs. 14 and 15 are not enough to show the differences of compared methods; thus, we calculate the AEMs in the second rows of Figs. 14 and 15

[10] Data link: http://www.rscloudmart.com/dataProduct/sample

[11] Note that, since our CS-Net and MRA-Net get weak performance according to the results on WorldView-2 and WorldView-3 data sets, hence, for the sake of brevity, we excluded these two methods from the analysis. Furthermore, for the same reason, we only show the results of the five CNN methods.

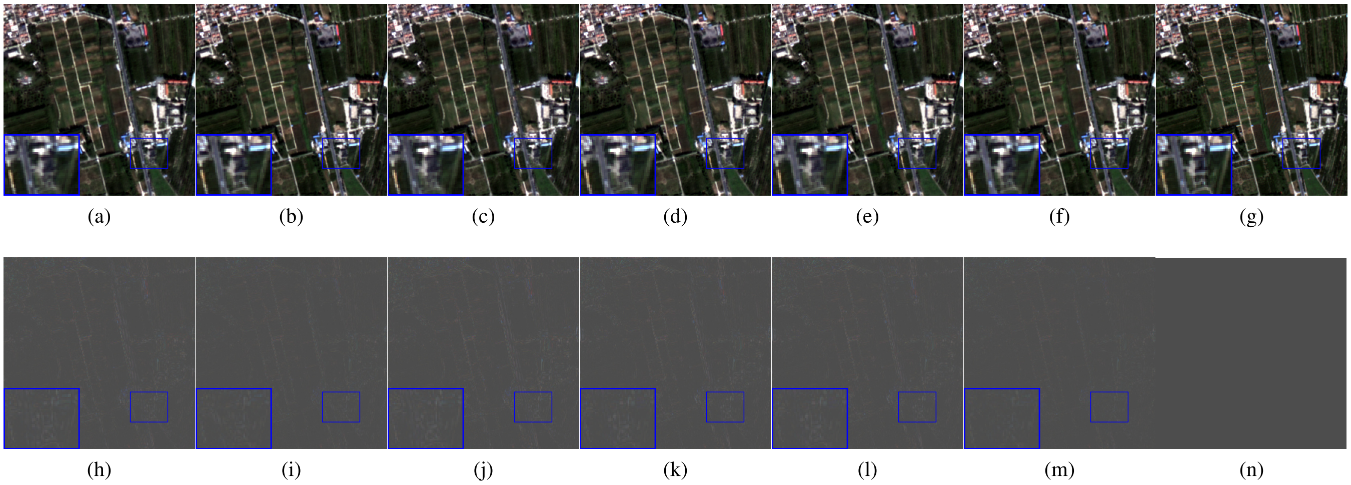


Fig. 14. Visual comparisons in natural colors of the most representative six approaches on the Guangzhou data set (sensor: GF-2). (First Row) Visual results. (Second Row) AEMs. (a) PNN. (b) DRPNN. (c) DiCNN1. (d) PanNet. (e) DMDNet. (f) Fusion-Net. (g) GT. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

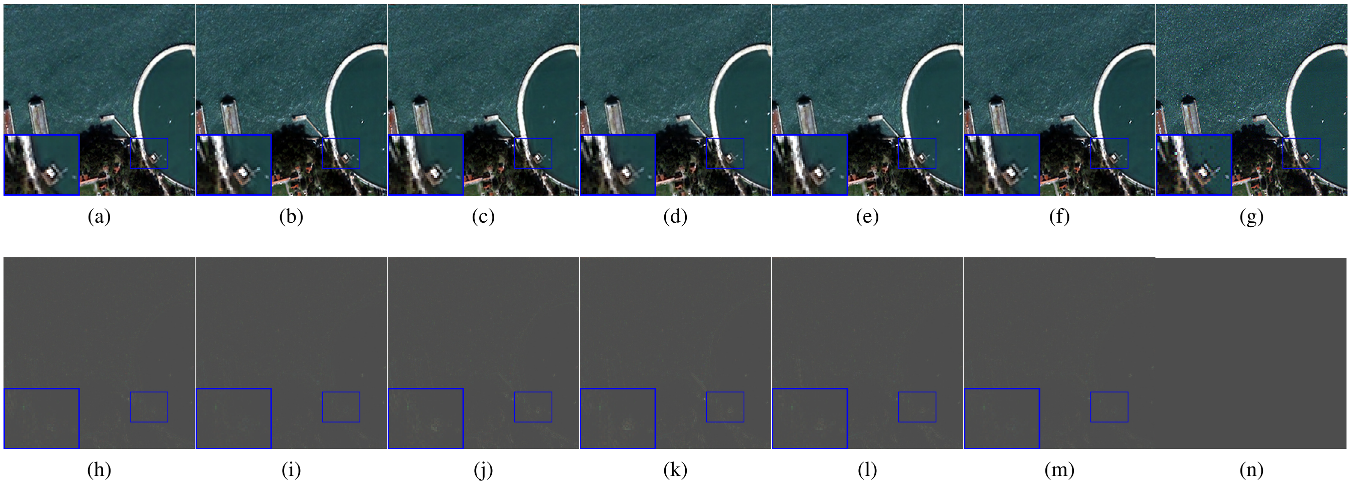


Fig. 15. Visual comparisons in natural colors of the most representative six approaches on the Indianapolis data set (sensor: QB). (First Row) Visual results. (Second Row) AEMs. (a) PNN. (b) DRPNN. (c) DiCNN1. (d) PanNet. (e) DMDNet. (f) Fusion-Net. (g) GT. (h) PNN. (i) DRPNN. (j) DiCNN1. (k) PanNet. (l) DMDNet. (m) Fusion-Net. (n) GT.

to aid the visual comparison. From the two figures, the proposed Fusion-Net clearly shows its spatial advantages getting lower image residuals (see the close-up boxes). Moreover, from Table VI, the proposed Fusion-Net still yields better quantitative assessments than the other compared approaches.

I. Discussion

Based on the previously shown results, it is clear that the CNN methods obtain better performance than the classical CS and MRA methods. This is mainly due to the fact that these methods exploit large-scale data for the training phase. In this section, we will discuss more about the detail images, the convergence, the network complexity, the computational times in both testing and training phases, the number of parameters (NoPs), and the giga floating-point operations per second (GFLOPs).

1) Detail Images: Unlike the previously shown AEMs, Fig. 16 displays the detailed images in order to point out the differences among the compared methods. The detailed images are obtained by taking the absolute value of the difference between the fused and the EXP images. From Fig. 16, the Fusion-Net gets the darker detail image, which demonstrates the effectiveness of the proposed method even exploiting this different representation of the fused outcomes.

2) Convergence: Fig. 17 exhibits the training errors of all the deep network methods with increasing iterations. It is worth to be noted that the maximum number of iterations for each method is the corresponding optimal iteration. It is straightforward that the training error of the proposed Fusion-Net (black line) reaches a lower level than those of the other approaches, which demonstrates that the Fusion-Net gets better training effectiveness.

3) Network Complexity: The proposed Fusion-Net is simpler than the PanNet. Comparing it with PanNet, Fusion-Net does not need to calculate the high-pass filtered version of the PAN image, thus reducing the training time with respect

TABLE VI

QUANTITATIVE ASSESSMENT OF THE COMPARED NETWORKS FOR THE GF-2 TESTING DATA SET (81 SAMPLES) AND THE QB TESTING DATA SET (48 SAMPLES). BEST RESULTS IN BOLDFACE

	SAM (± std)	**ERGAS** (± std)	**Q4** (± std)	**SCC** (± std)
Guangzhou (GF-2)				
PNN	1.6599 ± 0.3606	1.5707 ± 0.3243	0.9274 ± 0.0202	0.9281 ± 0.0206
DRPNN	1.4578 ± 0.2289	1.3735 ± 0.1876	0.9308 ± 0.0148	0.9384 ± 0.0052
DiCNN1	1.4948 ± 0.3814	1.3203 ± 0.3543	0.9445 ± 0.0211	0.9458 ± 0.0222
PanNet	1.3954 ± 0.3261	1.2239 ± 0.2828	0.9468 ± 0.0222	0.9558 ± 0.0123
DMDNet	1.2968 ± 0.3156	1.1281 ± 0.2669	0.9529 ± 0.0218	0.9644 ± 0.0100
Fusion-Net	**1.1795 ± 0.2714**	**1.0023 ± 0.2271**	**0.9627 ± 0.0167**	**0.9710 ± 0.0074**
Indianapolis dataset (QB)				
PNN	5.7993 ± 0.9474	5.5712 ± 0.4584	0.8572 ± 0.1481	0.9023 ± 0.0489
DRPNN	5.3667 ± 0.7721	5.270 ± 0.2809	0.8745 ± 0.1320	0.9177 ± 0.0454
DiCNN1	5.3071 ± 0.9957	5.231 ± 0.5411	0.8821 ± 0.1431	0.9224 ± 0.0506
PanNet	5.3144 ± 1.0175	5.1623 ± 0.6814	0.8833 ± 0.1398	0.9296 ± 0.0585
DMDNet	5.1197 ± 0.9399	4.7377 ± 0.6486	0.8907 ± 0.1464	0.9350 ± 0.0652
Fusion-Net	**4.5402 ± 0.7789**	**4.0508 ± 0.2666**	**0.9102 ± 0.1364**	**0.9547 ± 0.0457**

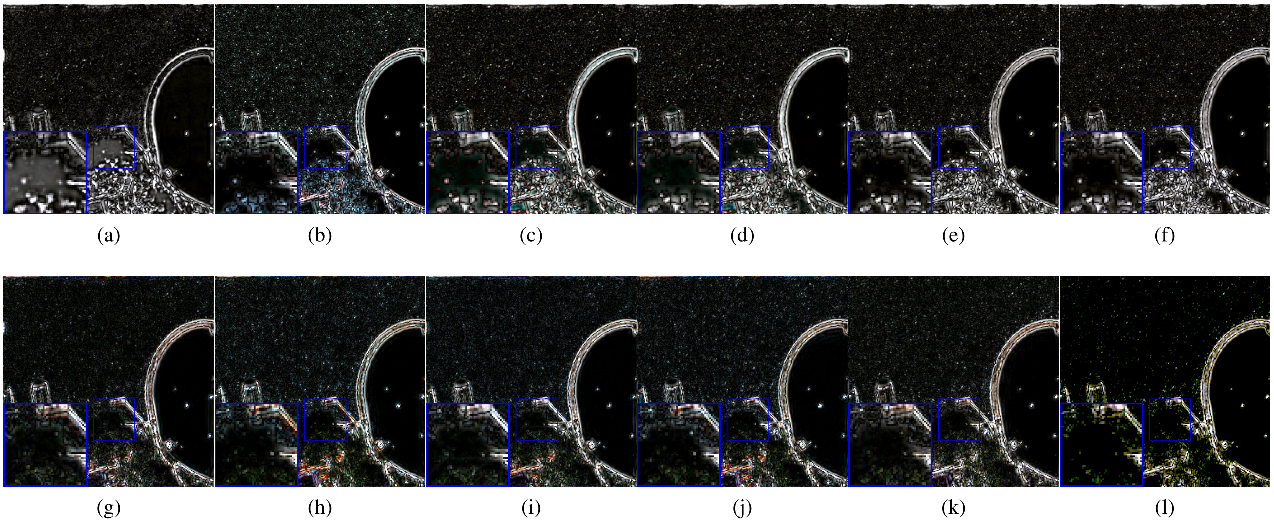


Fig. 16. Detail images of the different compared methods on a sample belonging to the Indianapolis data set (sensor: QB). (a) GS. (b) SFIM. (c) BDSD. (d) BDSD-PC. (e) GLP-Reg. (f) GLP-CBD. (g) PNN. (h) DRPNN. (i) DiCNN1. (j) PanNet. (k) DMDNet. (l) Fusion-Net.

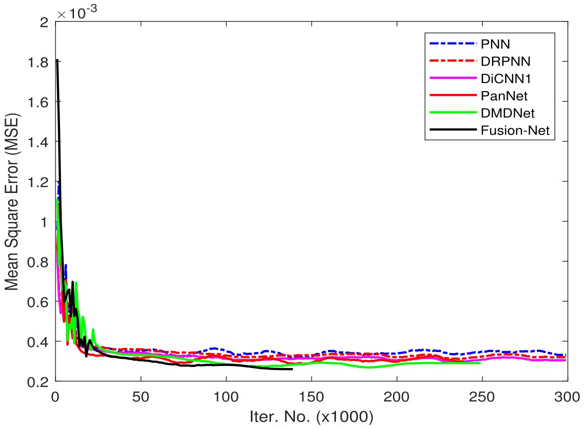


Fig. 17. Convergence curves for all the compared CNN methods on the WorldView-3 training data set. Note that we trained the PNN method with 1.12 × 10^6 iterations, but, here, we only show MSEs of the first 3 × 10^5 iterations for better display.

to PanNet. The architecture of DMDNet is similar to that of the PanNet, but DMDNet has a structure of grouped dilated convolution and, thus, is more complicated than PanNet. The architecture of DiCNN1 is slightly simpler than the PanNet and the Fusion-Net, but it is only a three-layer network meaning that is not easy to extract sufficient image features. The architecture of PNN is a simple three-layer network without any skip connection; thus, it is also not easy to extract enough image features from the simple network. In addition, the architecture of the DRPNN contains a skip connection and 11 layers, thus having a better feature extraction ability.

4) Testing Time: Table III reports the testing time of all the compared methods on two WorldView-3 data (i.e., Rio data set and Tripoli data set, both with size 256 × 256 × 8). Classical CS and MRA methods generally reach shorter testing times than that of the CNN methods. Furthermore, it is worth to be noted that CNN times are calculated on special hardware architecture (GPU); instead, to calculate the times of the CS and MRA approaches, a general-purpose CPU has been used. However, the testing time of the proposed networks can be considered acceptable on these data.

5) Training Time: The training times of all the CNNs are reported using the same training data set. The maximum

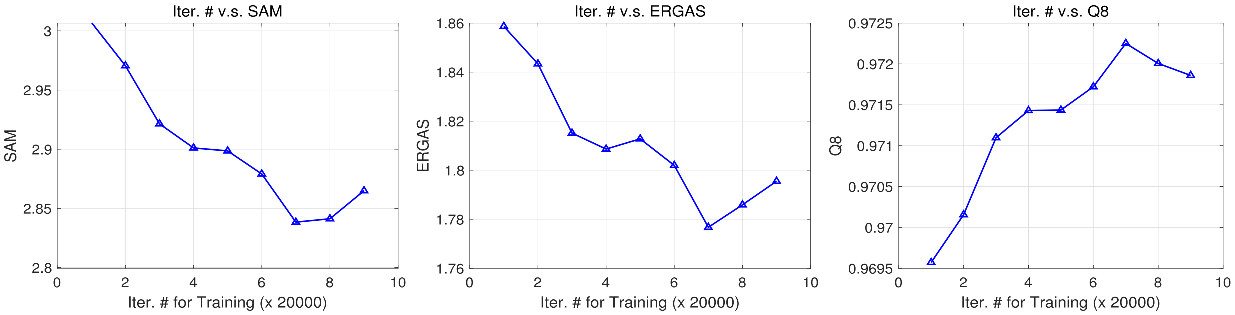


Fig. 18. Iteration number against quality metrics by averaging five runs on the Rio data set for the proposed Fusion-Net.

TABLE VII

Comparison of Training Times for All the Compared CNN Methods (Unit: Hours: Minutes)

PNN	DRPNN	DiCNN1	PanNet	DMDNet	Fusion-Net
1.12×10^6	3×10^5	3×10^5	2.4×10^5	2.5×10^5	1.4×10^5
25: 15	14: 25	7: 06	4: 32	5: 27	2: 21

TABLE VIII

Comparison of NoPs and GFLOPs for All the Compared CNN Methods

	PNN	DRPNN	DiCNN1	PanNet	DMDNet	Fusion-Net
NoPs	3.1×10^5	5.5×10^6	1.8×10^5	2.5×10^5	3.2×10^5	2.3×10^5
GFLOPs	0.427	7.619	0.192	0.340	0.359	0.323

iteration for each method is the optimal one used in the training phase. In Table VII, the proposed Fusion-Net yields the shortest training time mainly due to the fewer iterations when reaching convergence.

6) NoPs and GFLOPs: The NoPs and the GFLOPs of all the compared CNNs are reported in Table VIII. From Table VIII, it is clear that the DiCNN gets the best performance on the NoPs and the GFLOPs due to its simple architecture with only three convolutional layers. The proposed Fusion-Net holds the second place, which is better than other compared DL-based networks. The DRPNN approach gets the worse NoPs and GFLOPs since it involves more filters and the convolutional kernels with a larger size, i.e., 7 × 7.

7) Optimal Iteration Number for Fusion-Net: We want to investigate the optimal value of the iteration number for the proposed Fusion-Net. In order to select it, we consider an exemplary reduced resolution data set as the Rio data set. We calculated the performance metrics (the average of five runs) as in Fig. 18 taking the number of iterations that show the best overall quality. Thus, we refer to the value that gets the maximum Q8 index (around 140 000 iterations in Fig. 18) due to the fact that the Q8 can be considered an overall quality index. However, all the reduced resolution performance metrics are often in agreement with each other (see Fig. 18).

V. Conclusion

In this article, we investigated new architectures of CNNs for pansharpening. In particular, we focused our attention on DCNNs inspired by the classical fusion schemes exploited in CS and MRA methods. Thus, detail-based networks have been proposed and assessed on real WorldView-2, WorldView-3, GF-2, and QB data. The performance of the proposed ML methods has been compared with several state-of-the-art CS and MRA techniques and some powerful CNN-based methods for pansharpening. It has been demonstrated that the proposed Fusion-Net is able to get the best performance both at reduced and full resolutions. Finally, interesting features of the proposed Fusion-Net have been underlined from other points of view (e.g., computational burden, generalization capability, and robustness) comparing it with the other CNN-based methods.

References

[1] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," *IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5449–5457.

[2] L. He *et al.*, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.

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

[4] M. D. Mura, S. Prasad, F. Pacifici, P. Gamba, and J. Chanussot, "Challenges and opportunities of multimodality and data fusion in remote sensing," in *Proc. 22nd Eur. Signal Process. Conf. (EUSIPCO)*, 2014, pp. 106–110.

[5] 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.

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

[7] C. Souza, "Mapping forest degradation in the eastern Amazon from SPOT 4 through spectral mixture models," *Remote Sens. Environ.*, vol. 87, no. 4, pp. 494–506, Nov. 2003.

[8] C. Wu, B. Du, X. Cui, and L. Zhang, "A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion," *Remote Sens. Environ.*, vol. 199, pp. 241–255, Sep. 2017.

[9] 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.

[10] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6421–6433, Sep. 2019.

[11] 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.

[12] 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. 4, 2000.

[13] J. G. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3461–3472, Jan. 2000.

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

[15] M. J. Shensa, "The discrete wavelet transform: Wedding the a trous and Mallat algorithms," *IEEE Trans. Signal Process.*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

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

[17] 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, Jan. 2002.

[18] R. Restaino, G. Vivone, P. Addesso, and J. Chanussot, "A pansharpening approach based on multiple linear regression estimation of injection coefficients," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 1, pp. 102–106, Jan. 2020.

[19] G. Vivone, S. Marano, and J. Chanussot, "Pansharpening: Context-based generalized Laplacian pyramids by robust regression," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 9, pp. 6152–6167, Sep. 2020.

[20] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3418–3431, Jul. 2018.

[21] X. He, L. Condat, J. M. Bioucas-Dias, J. Chanussot, and J. Xia, "A new pansharpening method based on spatial and spectral sparsity priors," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4160–4174, Sep. 2014.

[22] Y. Jiang, X. Ding, D. Zeng, Y. Huang, and J. Paisley, "Pan-sharpening with a hyper-Laplacian penalty," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 540–548.

[23] T. Wang, F. Fang, F. Li, and G. Zhang, "High-quality Bayesian pansharpening," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 227–239, Jan. 2019.

[24] M. Moller, T. Wittman, and A. L. Bertozzi, "A variational approach to hyperspectral image fusion," *Proc. SPIE*, vol. 7334, pp. 73341E-1–73341E-10, Apr. 2009.

[25] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pansharpening," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2822–2834, Jul. 2013.

[26] J. Duran, A. Buades, B. Coll, and C. Sbert, "A nonlocal variational model for pansharpening image fusion," *SIAM J. Imag. Sci.*, vol. 7, no. 2, pp. 761–796, Jan. 2014.

[27] 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.

[28] H. A. Aly and G. Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Trans. Image Process.*, vol. 23, no. 6, pp. 2596–2608, Jun. 2014.

[29] C. Chen, Y. Li, W. Liu, and J. Huang, "SIRF: Simultaneous satellite image registration and fusion in a unified framework," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 4213–4224, Nov. 2015.

[30] G. Vivone *et al.*, "Pansharpening based on semiblind deconvolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 4, pp. 1997–2010, Apr. 2015.

[31] Q. Wei, N. Dobigeon, J.-Y. Tourneret, J. Bioucas-Dias, and S. Godsill, "R-FUSE: Robust fast fusion of multiband images based on solving a Sylvester equation," *IEEE Signal Process. Lett.*, vol. 23, no. 11, pp. 1632–1636, Nov. 2016.

[32] L.-J. Deng, G. Vivone, W. Guo, M. D. Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 535–539.

[33] Z.-Y. Zhang, T.-Z. Huang, L.-J. Deng, J. Huang, X.-L. Zhao, and C.-C. Zheng, "A framelet-based iterative pan-sharpening approach," *Remote Sens.*, vol. 10, no. 4, p. 622, Apr. 2018.

[34] L.-J. Deng, G. Vivone, W. Guo, M. D. Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4330–4344, Sep. 2018.

[35] G. Vivone, P. Addesso, R. Restaino, M. D. Mura, and J. Chanussot, "Pansharpening based on deconvolution for multiband filter estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 540–553, Jan. 2019.

[36] L.-J. Deng, M. Feng, and X.-C. Tai, "The fusion of panchromatic and multispectral remote sensing images via tensor-based sparse modeling and hyper-Laplacian prior," *Inf. Fusion*, vol. 52, pp. 76–89, Dec. 2019.

[37] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[38] 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.

[39] M. R. Vicinanza, R. Restaino, G. Vivone, M. D. Mura, and J. Chanussot, "A pansharpening method based on the sparse representation of injected details," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 1, pp. 180–184, Jan. 2015.

[40] 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.

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

[42] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," in *Proc. Int. Workshop Remote Sens. Intell. Process. (RSIP)*, May 2017, pp. 1–4.

[43] N. Li, N. Huang, and L. Xiao, "PAN-sharpening via residual deep learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 5133–5136.

[44] 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.

[45] F. Luo, B. Du, L. Zhang, L. Zhang, and D. Tao, "Feature learning using spatial–spectral hypergraph discriminant analysis for hyperspectral image," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2406–2419, Jul. 2019.

[46] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral–spatial unified networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 10, pp. 5893–5909, Oct. 2018.

[47] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, and Y. Zhang, "Residual encoder–decoder conditional generative adversarial network for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 9, pp. 1573–1577, Sep. 2019.

[48] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[49] S. Eghbalian and H. Ghassemian, "Multi spectral image fusion by deep convolutional neural network and new spectral loss function," *Int. J. Remote Sens.*, vol. 39, no. 12, pp. 3983–4002, Jun. 2018.

[50] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.

[51] 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.

[52] L. Liu *et al.*, "Shallow–deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1772–1783, 2020.

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

[54] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "HPGAN: Hyper-spectral pansharpening using 3-D generative adversarial networks," *IEEE Trans. Geosci. Remote Sens.*, early access, May 20, 2020, doi: 10.1109/TGRS.2020.2994238.

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

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[57] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: Assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.*, vol. 63, pp. 691–699, Jun. 1997.

[58] 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, no. 5, pp. 591–596, May 2006.

[59] 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.

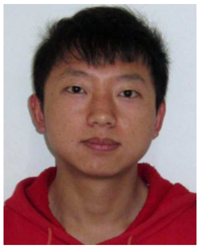
[60] X. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, “Deep multiscale detail networks for multiband spectral image sharpening,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 2, 2020, doi: 10.1109/TNNLS.2020.2996498.

[61] 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. JPL Airborne Geosci. Workshop; AVIRIS Workshop*, Pasadena, CA, USA, 1992, pp. 147–149.

[62] L. Wald, “Data fusion: Definitions and architectures: Fusion of images of different spatial resolutions,” in *Presses des MINES*. Paris, France: Ecole des Mines, 2002.

[63] 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, Jan. 1998.

[64] A. Garzelli and F. Nencini, “Hypercomplex quality assessment of multi/hyperspectral images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.



Liang-Jian Deng (Member, IEEE) received the B.S. and Ph.D. degrees in applied mathematics from the School of Mathematical Sciences, University of Electronic Science and Technology of China (UESTC), Chengdu, China, in 2010 and 2016, respectively.

From 2013 to 2014, he was a Joint-Training Ph.D. Student with Case Western Reserve University, Cleveland, OH, USA. From 2017 to 2018, he was a Post-Doctorate with Hong Kong Baptist University (HKBU), Hong Kong. In addition, he also stayed with the Isaac Newton Institute for Mathematical Sciences (Cambridge University), Cambridge, U.K., and HKBU for short visits. He is an Associate Professor with the School of Mathematical Sciences, UESTC. His research interests include the use of partial differential equations (PDE), optimization modeling, and deep learning to address several tasks in image processing and computer vision, e.g., resolution enhancement, restoration, etc.

Gemine Vivone (Senior Member, IEEE) received the B.Sc. *(summa cum laude)*, M.Sc. *(summa cum laude)*, and Ph.D. degrees in information engineering from the University of Salerno, Salerno, Italy, in 2008, 2011, and 2014, respectively.

In 2013, he was a Visiting Scholar with the Grenoble Institute of Technology (INPG), Grenoble, France. In 2014, he joined the North Atlantic Treaty Organization (NATO) Science and Technology Organization (STO), Centre for Maritime Research and Experimentation (CMRE), La Spezia, Italy as a Scientist. In 2019, he was an Assistant Professor with the University of Salerno. He is the Leader of the Image and Signal Processing Working Group of the IEEE Image Analysis and Data Fusion Technical Committee. He is also a Researcher with the National Research Council, Rome, Italy, and an Adjunct Professor with the University of Salerno. His main research interests focus on statistical signal processing, detection of remotely sensed images, data fusion, and tracking algorithms.

Dr. Vivone is an Editorial Board Member for Multidisciplinary Digital Publishing Institute (MDPI) *Remote Sensing*, MDPI *Sensors*, and MDPI *Encyclopedia*. He received the Symposium Best Paper Award at the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) in 2015 and the Best Reviewer Award of the IEEE Transactions on Geoscience and Remote Sensing (GRS) in 2017. He served as a Guest Associate Editor for several special issues. He is an Associate Editor for the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS (GRSL).

Cheng Jin is pursuing the B.S. degree with the School of Optoelectronic Science and Technology, University of Electronic Science and Technology of China (UESTC), Chengdu, China.

His research interests include digital image processing utilizing deep learning, e.g., resolution enhancement, etc.

Jocelyn Chanussot (Fellow, IEEE) 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 Université de Savoie, Annecy, France, in 1998.

Since 1999, he has been with Grenoble INP. He was a Visiting Scholar with Stanford University, Stanford, CA, USA, KTH, Stockholm, Sweden, and National University of Singapore (NUS), Singapore. Since 2013, he has been an Adjunct Professor with the University of Iceland, Reykjavik, Iceland. From 2015 to 2017, he was a Visiting Professor with the University of California at Los Angeles (UCLA), Los Angeles, CA, USA. He holds the AXA Chair in Remote Sensing and is an Adjunct Professor with the Chinese Academy of Sciences, Aerospace Information Research Institute, Beijing, China. He is a Professor of signal and image processing with Grenoble INP. His research interests include image analysis, hyperspectral remote sensing, data fusion, machine learning, and artificial intelligence.

Dr. Chanussot was a member of the Machine Learning for Signal Processing Technical Committee of the IEEE Signal Processing Society from 2006 to 2008, and the Institut Universitaire de France from 2012 to 2017. He is the Founding President of the IEEE Geoscience and Remote Sensing (GRS) French chapter from 2007 to 2010 which received the 2010 IEEE Geoscience and Remote Sensing (GRS)-S Chapter Excellence Award. He has received multiple outstanding paper awards. He was the Vice-President of the IEEE GRS Society (GRSS), the In-Charge of meetings and symposia from 2017 to 2019. 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 from 2009 to 2011 and the Co-Chair of the GRS Data Fusion Technical Committee from 2005 to 2008. He was the Program Chair of the IEEE International Workshop on Machine Learning for Signal Processing in 2009. He is an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, and the PROCEEDINGS OF THE IEEE. He was the Editor-in-Chief of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING from 2011 to 2015. In 2014 he served as a Guest Editor for the *IEEE Signal Processing Magazine*. He is a Highly Cited Researcher (Clarivate Analytics/Thomson Reuters from 2018 to 2019.