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 data with simultaneously acquired multispectral data with lower spatial resolution is a hot topic, which is often called pansharpening. In this paper, we exploit the combination of machine learning techniques and fusion schemes introduced to address the pansharpening problem. In particular, deep convolutional neural networks are proposed to solve this issue. These latter are combined first with the traditional component substitution and multi-resolution analysis fusion schemes in order to estimate the non-linear injection models that rule the combination of the upsampled low resolution multispectral image with the extracted details exploiting the two philosophies. Furthermore, inspired by these two approaches, we also developed another deep convolutional neural network for pansharpening. This is fed by the direct difference between the panchromatic image and the upsampled low resolution multispectral 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*—Deep Convolutional Neural Network, Component Substitution, Multi-resolution Analysis, Pansharpening, Image Fusion, Remote Sensing.**

I. Introduction

Pansharpening has become a fundamental problem in remote sensing image processing, since it can fuse a *high* 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 papers [5], [6]. Furthermore, pansharpened products have attracted the interest of some commercial companies, *e.g.*, Google Earth, as well as 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, multi-resolution 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

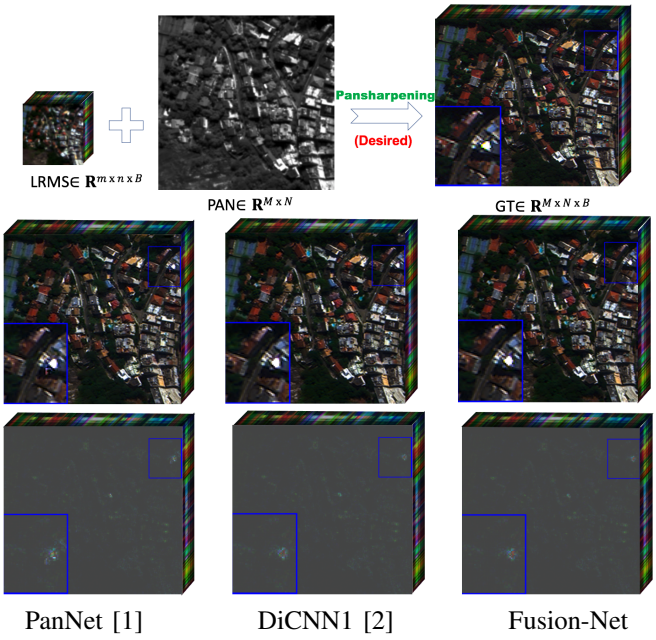


Fig. 1: First row: flowchart for pansharpening on an 8-band WorldView-3 satellite data with a spatial resolution factor equal to 4. The figure includes the low spatial resolution MS image, the PAN image, and the ground-truth image. Second and third rows: the pansharpened images and the corresponding absolute error maps 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 absolute error map implying superior performance with respect to the competitors.

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

G. Vivone is with the Department of Information Engineering, Electrical Engineering and Applied Mathematics, University of Salerno, 84084 Fisciano, Italy and with the Institute of Methodologies for Environmental Analysis, CNR-IMAA, 85050 Tito Scalo, Italy. (e-mails: gvivone@unisa.it; gemine.vivone@imaa.cnr.it).

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

J. Chanussot is with Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, Grenoble, 38000, France (e-mail: jocelyn.chanussot@gipsa-lab.grenoble-inp.fr).

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 intensity component. Then, the (possibly equalized) PAN image can be substituted with the intensity component. The sharpened version of the MS image is obtained thanks to the inverse projection bringing the data in the original multispectral 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 with local parameter estimation (BDSD) [9], the robust band-dependent spatial-detail (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 a multi-resolution analysis 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. 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: Bayesian methods [21]–[23], 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 an 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 *e.g.*, [1], [2], [40]–[54]. The CNN-based methods depend on the large-scale dataset training to learn a non-linear functional mapping between the low spatial resolution MS images and the high spatial resolution multispectral images. After the training phase, it is easy to predict/compute the pansharpened image by the learned non-linear mapping. In [41], Masi *et al.* [41] proposed first 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. In [52], Liu *et al.* 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. In [1], Yang *et al.* 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 preservation, 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. In [2], He *et al.* 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 detail 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 paper, we propose deep CNNs to address the pansharpening problem, even accounting for fusion schemes proposed in literature. In particular, we focus our attention on traditional CS and MRA frameworks. The details are extracted using these two philosophies. Instead, the non-linear injection model is estimated through CNNs. These approaches are here named CS-Net and MRA-Net, respectively. Inspired by these solutions, we further investigate on 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 pre-processing step using detail 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 non-linear and local injection model. This approach will be called Fusion-Net from hereon. The proposed approaches are tested on several datasets acquired by WorldView-2, WorldView-3, GaoFen (GF)-2 and QuickBird (QB) datasets. The experimental analysis is conducted both at reduced and full resolutions. The benchmark consists of state-of-the-art CS and MRA approaches and machine learning 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.

In summary, the main contributions of this work are:

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 reaching state-of-the-art performance, see, *e.g.*, the comparison among the high performance CNNs in Fig. 1 for a WorldView-3 dataset. Moreover, the Fusion-Net has a simple architecture with fewer network parameters, thus resulting more effective than some previously developed network architectures for pansharpening.
3) A broad experimental analysis has been provided based

on several datasets. 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 datasets has been provided to the readers for all the considered CNN approaches.

The paper is organized as follows. The related works and motivations are introduced in Sect. II. The proposed three network architectures will be detailed in Sect. III. Sect. IV is devoted to the description of the experimental results and the related discussions. Finally, conclusions are drawn in Sect. 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 firstly introduce them in this section, then we will move towards the motivations under the choice of the proposed network architectures.

A. CS

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

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

where $\widehat{\mathbf{MS}}_i \in \mathbb{R}^{M \times N}$ is the i-th band of the high spatial resolution MS image, $\widetilde{\mathbf{MS}}_i \in \mathbb{R}^{M \times N}$ is the i-th band of the upsampled version of the low spatial resolution MS image, $g_i \in \mathbb{R}$ is the i-th 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 i-th 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, *e.g.*, [9], [11]–[13].

Equation (1) could be further rewritten in the following multi-band 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 $\widehat{\mathbf{MS}}_i$, $i = 1, 2, \cdots, B$ and $\widetilde{\mathbf{MS}}_i$, $i = 1, 2, \cdots, B$, 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, $\mathbf{g} = (g_1, g_2, \cdots, g_B)^T \in \mathbb{R}^B$ is a vector of coefficients g_i as in (1), and $\odot$ is an operator indicating that the i-th element of $\mathbf{g}$ multiplies the i-th spectral band of $\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}}$.

B. MRA

Similar as 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 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, *e.g.*, [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 *non-linear* approaches, aiming of replacing the detail injection phases in both CS and MRA methods. Deep convolutional neural networks (DCNNs) can easily manage this non-linear 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: i) detail extraction and ii) detail injection into the original MS image. But, we address the issue of non-linear 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. Afterwards, 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 $\mathbf{g} \odot \left(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}} \right)$. In this equation, the upsampled MS image $\widetilde{\mathbf{MS}}$ holds the spatial information at low resolution and $\left(\mathbf{P}^{\mathbf{D}} - \mathbf{I}_L^{\mathbf{D}} \right)$ provides the high frequency details, injected through $\mathbf{g}$.

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

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

where $f_{\mathbf{\Theta}_{CS}}$ is the non-linear mapping with the network parameter $\mathbf{\Theta}_{CS}$ that could be learned from a large-scale training dataset. Several solutions to get the weights $\omega_i, i = 1, \cdots, B$

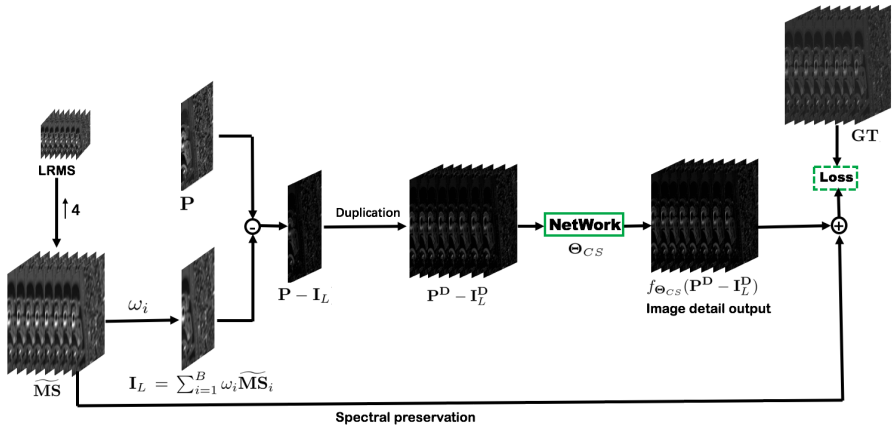


Fig. 2: The architecture of the CS-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork", please refer to Sect. III-D.

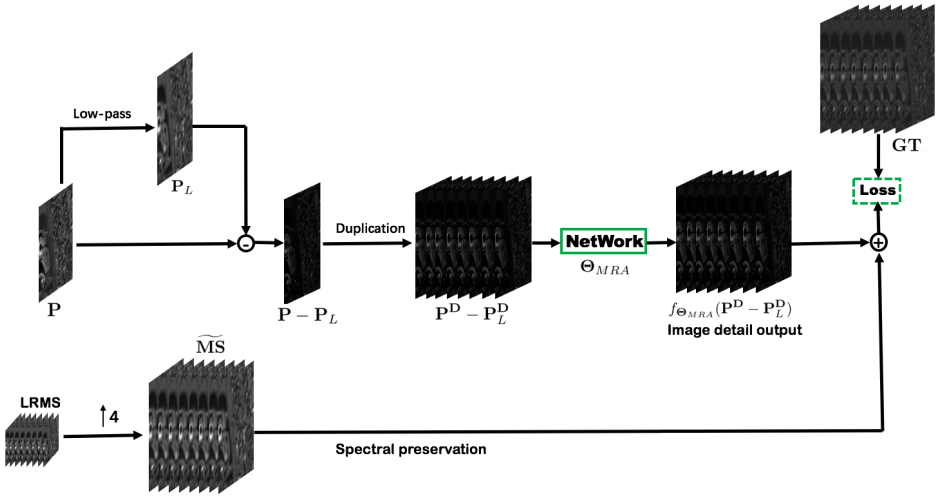


Fig. 3: The architecture of the MRA-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork", please refer to Sect. III-D.

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 built 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:

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

where n represents the number of training examples, $\|\cdot\|_F$ is the Frobenius norm, and $\mathbf{GT}_{\{k\}}$ is the k-th 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 $\mathbf{\Theta}_{CS}$. 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 $\mathbf{g}\odot\left(\mathbf{P}^{\mathbf{D}}-\mathbf{P}_L^{\mathbf{D}}\right)$. Again, we ignore the injection coefficients $\mathbf{g}$ by imposing a non-linear mapping function estimated through a DCNN fed by $\left(\mathbf{P}^{\mathbf{D}}-\mathbf{P}_L^{\mathbf{D}}\right)$. Therefore, the MRA-Net can be summarized as follows:

$$\widehat{\mathbf{MS}} = \widetilde{\mathbf{MS}} + f_{\mathbf{\Theta}_{MRA}}(\mathbf{P}^{\mathbf{D}}-\mathbf{P}_L^{\mathbf{D}}), \quad (6)$$

where $f_{\mathbf{\Theta}_{MRA}}$ is the non-linear mapping function with network parameters $\mathbf{\Theta}_{MRA}$. 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

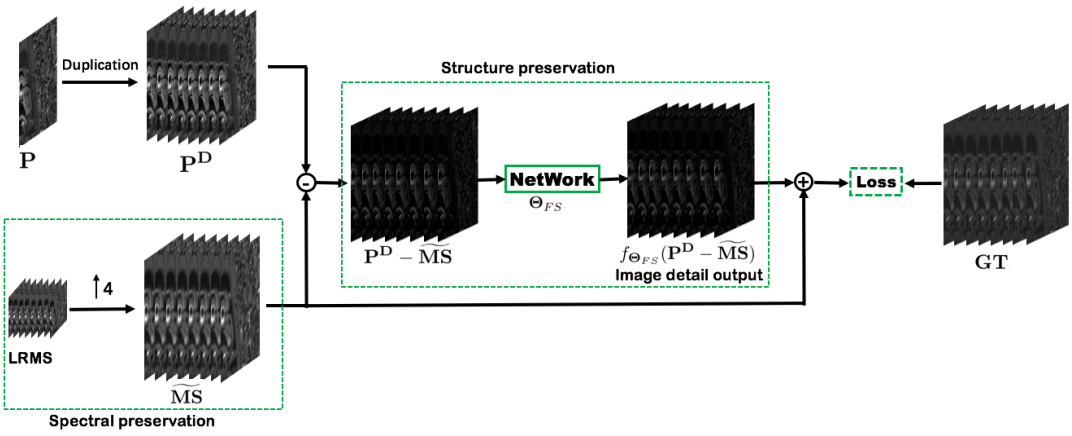


Fig. 4: The architecture of the Fusion-Net. The upsampling is performed using a polynomial kernel with 23 coefficients [17]. For "NetWork", please refer to Sect. III-D.

the MRA-Net is defined as follows:

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

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, $\mathbf{P^D}$, 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 $\mathbf{I}_L^{\mathbf{D}}$ or $\mathbf{P}_L^{\mathbf{D}}$. 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 $\mathbf{P^D} - \widetilde{\mathbf{MS}}$ 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.

Similarly 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 non-linear injection model. The Fusion-Net can be summarized as follows:

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

where $f_{\mathbf{\Theta}_{FS}}$ is the non-linear mapping with network parameters $\mathbf{\Theta}_{FS}$.

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:

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

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, thus improving performance of existing methods. Please, refer to Sect. 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 literature, called ResNet [56], as the subnetwork of the proposed architectures, since the ResNet can bring the conventional CNN to deeper layers leading to 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

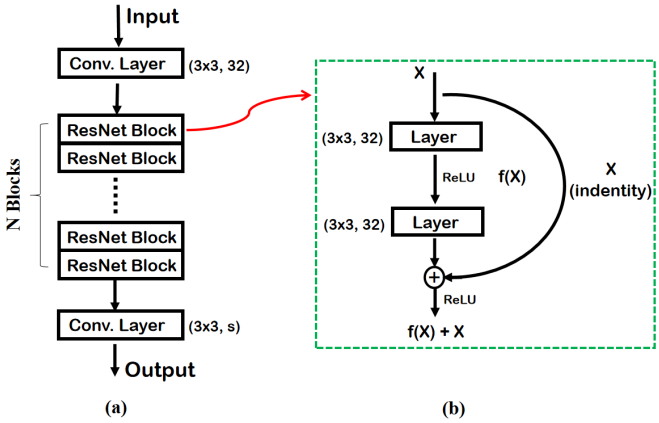


Fig. 5: (a) The structure of "NetWork" with several ResNet blocks, see the solid green boxes in Figs. 2-4. Note that $(3 \times 3, 32)$ represents 32 convolution kernels with size 3×3, and s depends on the number of multispectral bands (*e.g.,* for 4-band image $s = 4$ and for 8-band image $s = 8$). (b) The details of one ResNet block [56] that is used in our architectures. Each ResNet block contains two non-linear 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 Tab. I.

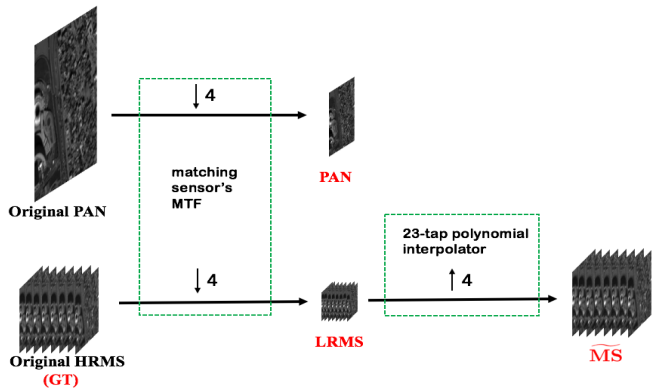


Fig. 6: The generation process of the training dataset by Wald's protocol. Note that the data indicated with the red text are the generated training data used to feed the networks, *i.e.*, the GT, the low spatial resolution MS (LRMS) image, the PAN, and the upsampled MS image ($\widetilde{\mathbf{MS}}$).

control the final convolutional layers, aiming to achieve the best performance (see the parameter setting in Sect. IV).

E. Generation of training data

In this work, we train the CNNs on WorldView-3 (8-bands) satellite datasets that can be easily downloaded on the public website[1]. After downloading the datasets, 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, then splitting them into 70/20/10% for training (8,806 examples[2])/validation (2,516 examples)/testing (1,258 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 dataset by Wald's protocol is illustrated in Fig. 6. It mainly contains the following steps: i) 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*; ii) Taking the original MS image as *the training GT image*; iii) 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 i)-iii), it is easy to generate the training data. The validation and testing datasets 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 of all, the employed sensors, the benchmark, and the adopted quality indexes will be described. Afterwards, the experimental analysis both at reduced and full resolutions will be described.

A. Datasets

Several datasets 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 m 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 4-band (red, green, blue, and near-infrared) datasets. In particular, QuickBird (QB) data are considered having a spatial resolution of 2.4 m and 0.61 m for the MS and PAN images, respectively, and a radiometric resolution of 11 bits. Finally, images acquired by the GaoFen (GF)-2 sensor have been exploited with a spatial resolution of 3.2 m and 0.8 m for the MS and PAN images, respectively, and a radiometric resolution of 10 bits (please, see Sect. IV-H for more details).

B. Benchmark

The proposed benchmark consists of the following methods: the MS image interpolation using a polynomial kernel with

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

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

23 coefficients (EXP) [17], the Gram-Schmidt sharpening approach (GS) [12], the smoothing filter-based intensity modulation (SFIM) [13], the partial replacement adaptive component substitution approach (PRACS) [11], the band-dependent spatial-detail method (BDSD) [9], the robust band-dependent spatial-detail approach (BDSD-PC) [10], the GLP with MTF-matched filter [58] and multiplicative injection model [59] (GLP-HPM), the GLP with MTF-matched filter [58] and regression-based injection model (GLP-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], the state-of-the-art CNN-based method for pansharpening (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 8GB.

C. Quality assessment

The performance assessment is conducted both at reduced and at 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 4-band images (Q4) and 8-band images (Q8) [64]. In particular, the ideal value for Q4, Q8 and SCC is 1, while for SAM and ERGAS is 0. Furthermore, to evaluate the performance at full resolution, we employ the QNR, the D_λ, and the 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 Sect. III-E, the training data for PanNet and DiCNN1 in this work are different from that of their original papers, 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 Tab. I[9].

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 dataset. Thus, this analysis is only devoted to compare CNN-based approaches trained on the same examples.

In Tab. II, we show first the average quantitative results of the different methods on the testing dataset 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, $16\times16\times8$ for the original low spatial resolution MS image, and $64 \times 64 \times 8$ for the GT image. From Tab. 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 demonstrate the robustness of the proposed Fusion-Net. In particular, having a look at Tab. II, 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 Sect. 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 datasets 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 Sect. III-E. The two datasets 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. Tab. 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 Tripoli dataset is used.

The visual analysis further corroborates these numerical results. Indeed, in Fig. 7 (Rio dataset), 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 the six CNN methods perform significantly better than the classical methods (both spatially and spectrally).

[3]Note that the given source code in Open Remote Sensing does not contain the trained models for WV2 and WV3, thus we re-implemented the network with default parameters in Python using Tensorflow for simplicity of comparison.

[4]It is not easy to find the source code, thus we re-implemented 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.

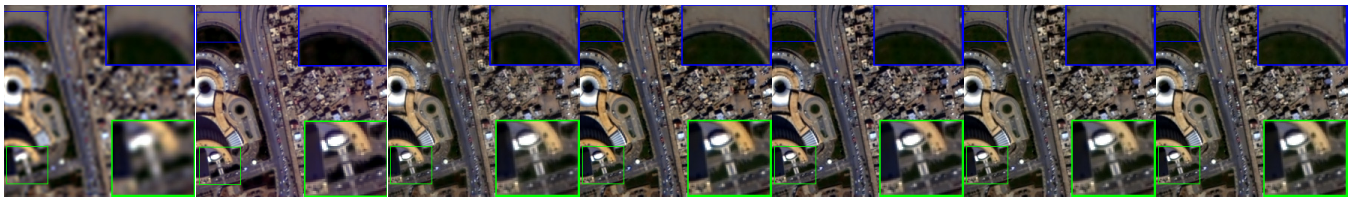
[7]DMDNet has been implemented by ourselves.

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

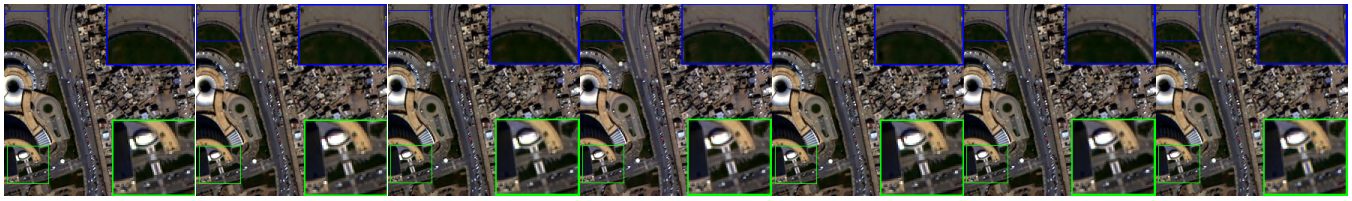
[9]Note that PanNet, CS-Net, MRA-Net, and Fusion-Net use ResNet blocks. If the number of layers for one of these networks is 10, that means that there are 4 ResNet blocks (each block with two layers) and two extra input and output layers.

TABLE I: Optimal parameters for the compared deep convolutional neural networks. Notation: **Iter. #** (iteration number), **Bs** (mini-batch 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



(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. 7: Visual comparisons in natural colors of the most representative 13 approaches on Rio dataset (WorldView-3).

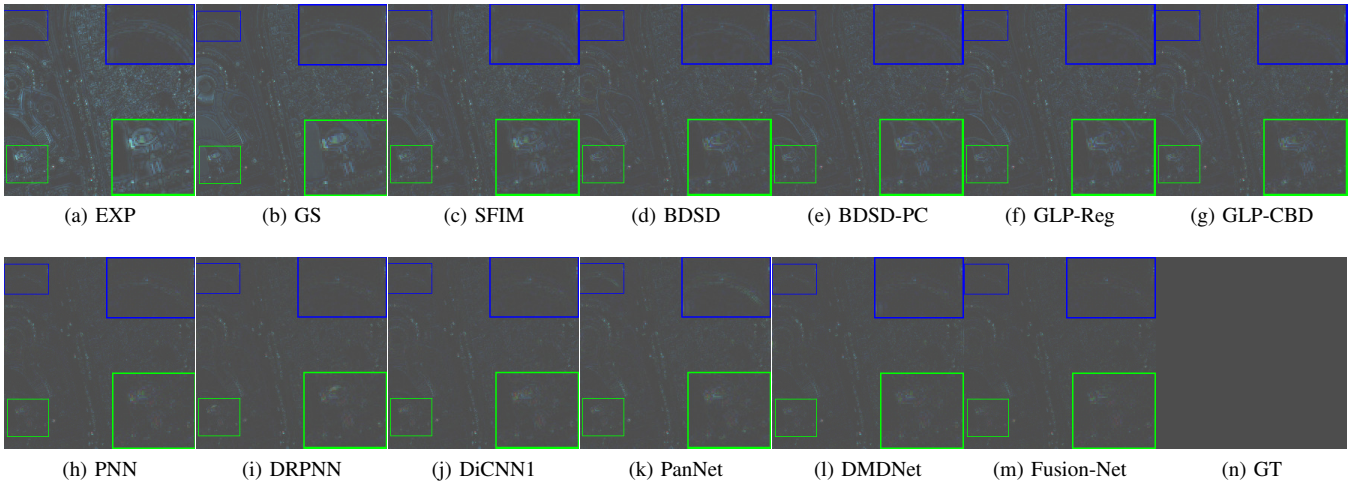
TABLE II: Quantitative comparison of the compared deep networks for the testing dataset that includes 1258 samples. Best results in boldface.

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

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-bits RGB images instead of 11-bits 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 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 Tripoli dataset in Figs. 9-10.

F. Full resolution assessment

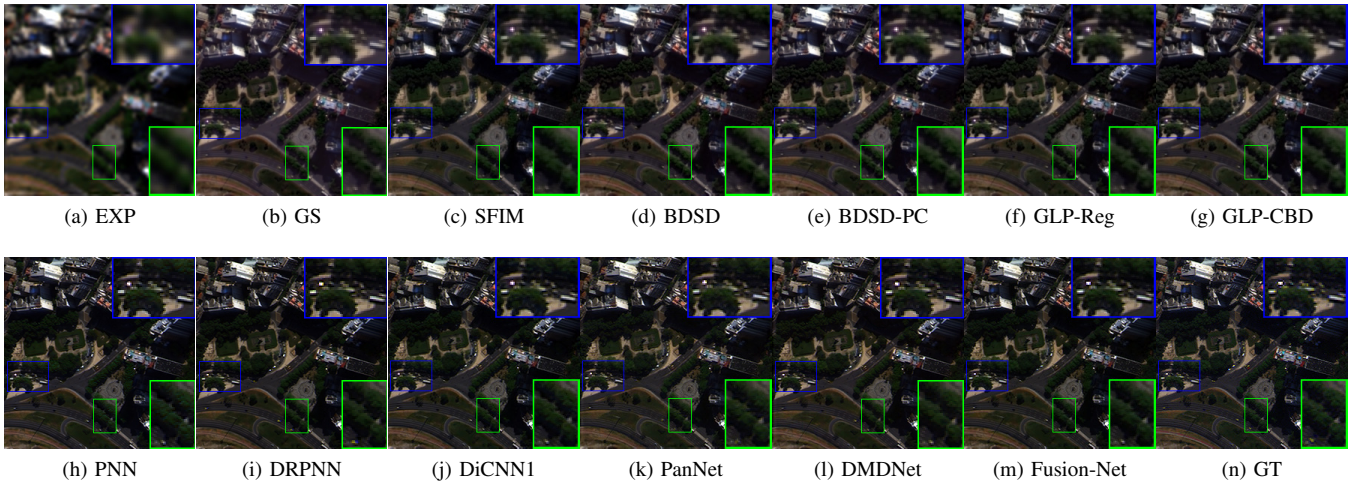
In this section, we test the performance of the proposed benchmark at the original (full) scale. In this case, the GT image is not available requiring quality indexes without reference for performance assessment purposes. We exploited 30 image pairs (MS and PAN) of WorldView-3 data at the original scale for testing the approaches using the QNR as quality index. Tab. 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, 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 dataset, here named Tripoli-OS dataset. 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



(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: Absolute error maps 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

Fig. 9: Visual comparisons in natural colors of the most representative 13 approaches on Tripoli dataset (WorldView-3).

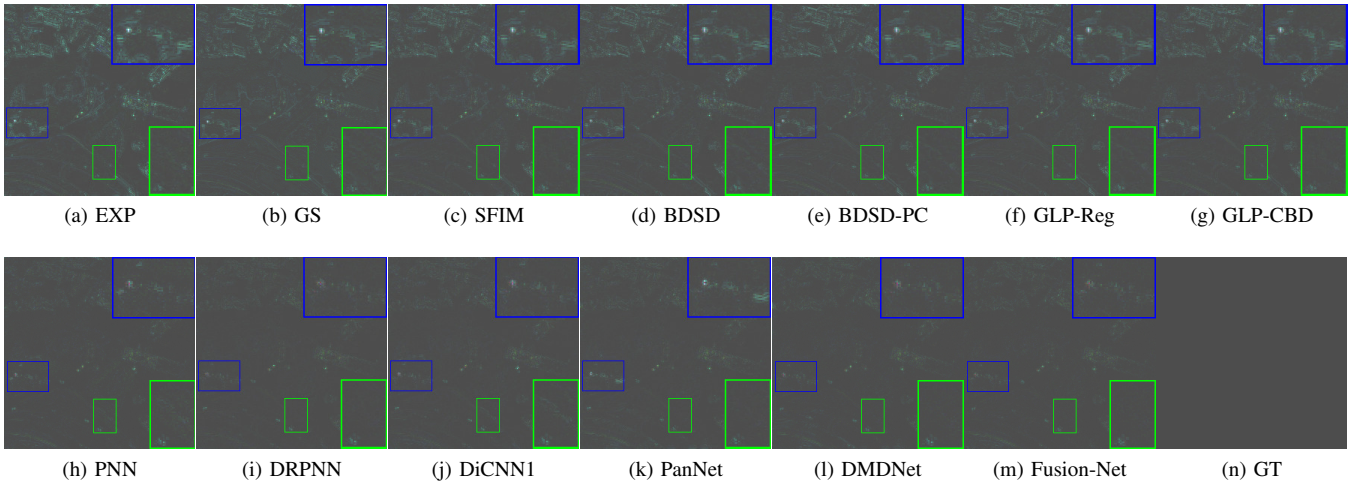
GLP-CBD), whereas all the CNN-based methods significantly outperform the classical approaches, both spatially and spectrally. Furthermore, the proposed Fusion-Net obtains a better spatial performance than that of the other five CNN-based methods. In the meanwhile, Fusion-Net is also able to preserve 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 fusing data acquired by different sensors. To this aim, we exploit another dataset acquired by another 8-bands 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 dataset acquired by the WorldView2 sensor. Quantitative results reported in Tab. 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 4-band datasets

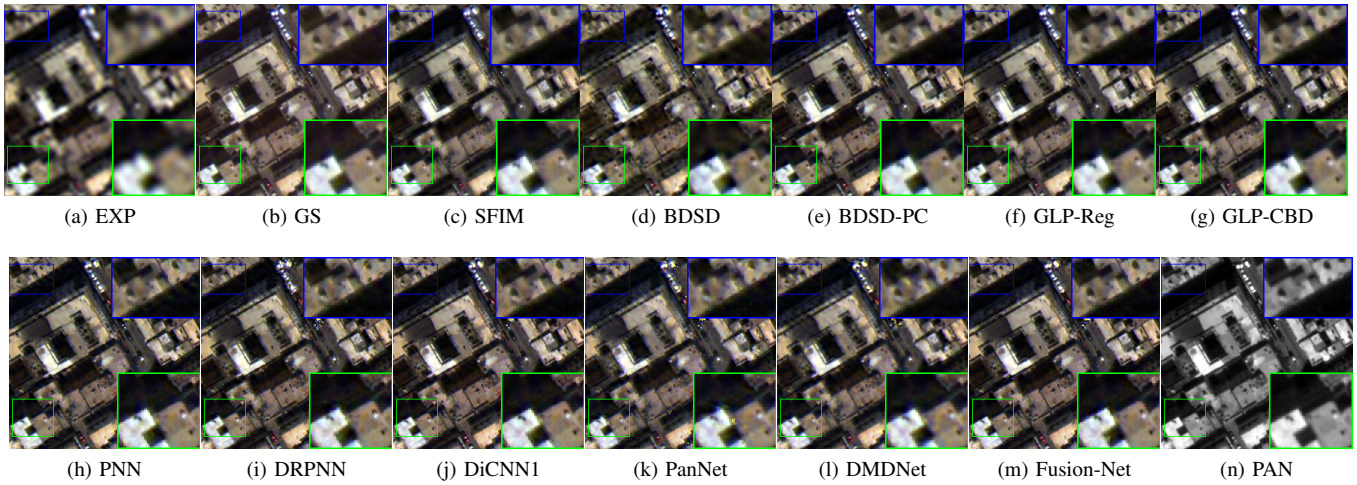
In this section, we will extend the performance assessment to 4-band datasets, *i.e.*, acquired by the GF-2 and the QB sensors.



(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: Absolute error maps 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) PAN

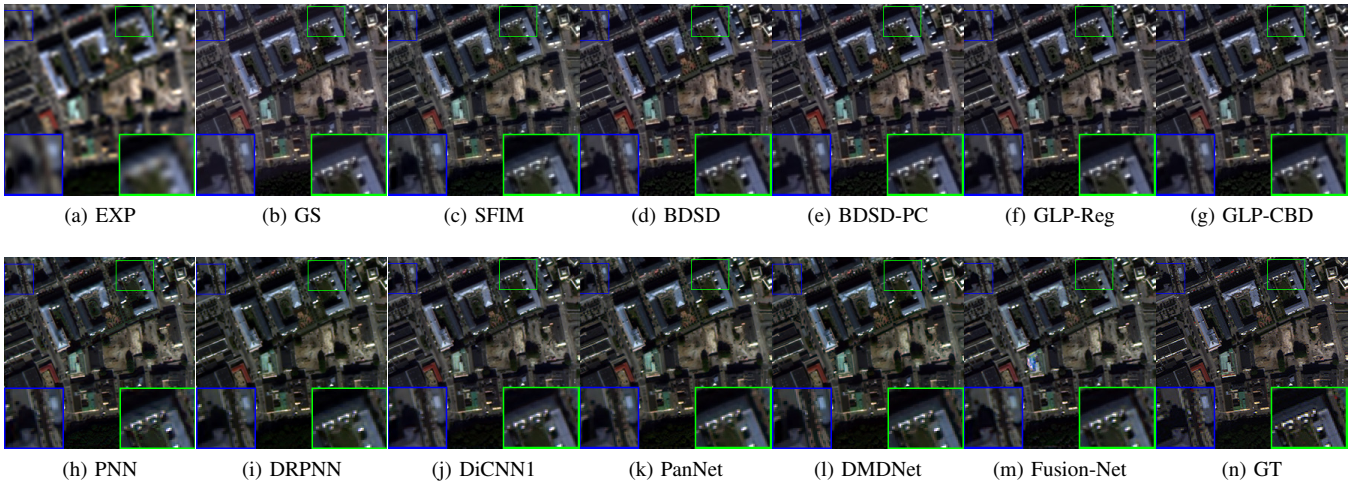
Fig. 11: Visual comparisons in natural colors of the most representative 13 approaches on Tripoli-OS dataset (WorldView-3) at the original scale.

About the data simulation, we also follow the way described in Sect. III-E to generate the training and testing data. For the QB test case, we downloaded a large dataset ($4906 \times 4906 \times 4$) acquired over the city of Indianapolis cutting it into two parts. The left part ($4906 \times 3906 \times 4$) is used to simulate 20685 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 dataset ($6907 \times 7300 \times 4$) over the city of Beijing from the website [10] to simulate 21607 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: $256 \times 256 \times 4$).

Fig. 14 and Fig. 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 Fig. 14 and Fig. 15 are not enough to show the differences of compared methods, thus we calculate the absolute error maps (AEMs) in the second rows of Fig. 14 and Fig. 15 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 Tab. VI, the proposed Fusion-Net still yields better quantitative assessments than the other compared approaches.

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

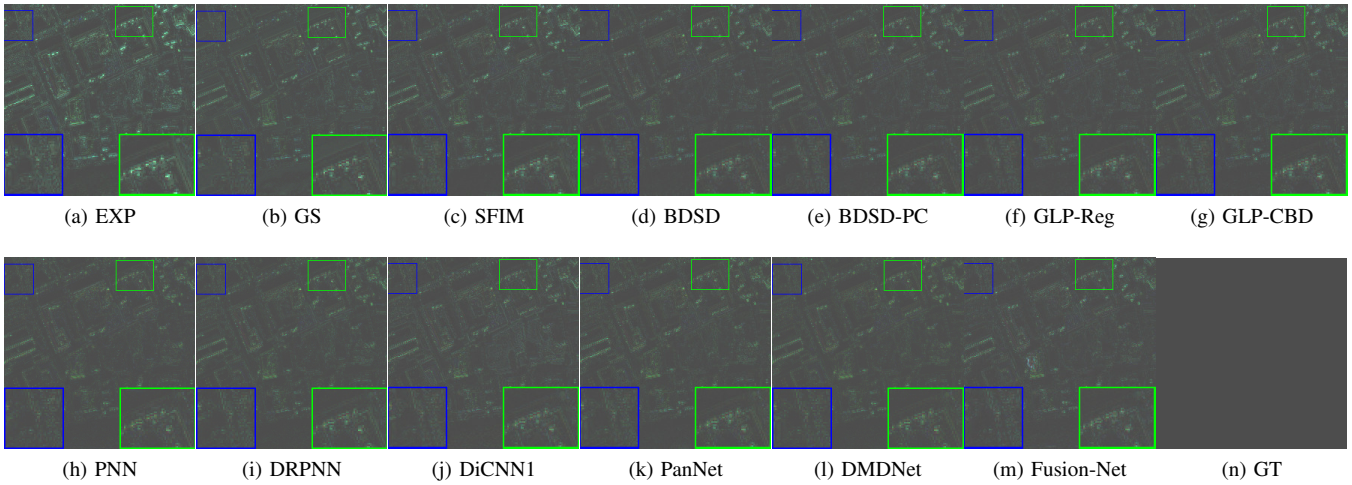
[11]Note that, since our CS-Net and MRA-Net get weak performance according to the results on WorldView-2 and WorldView-3 datasets. Hence, for the sake of brevity, we excluded these two methods from the analysis. Furthermore, for the same reason, we only show the results of the five CNN methods.



(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. 12: Visual comparisons in natural colors of the most representative 13 approaches on Stockholm dataset (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: Absolute error maps of Fig. 12.

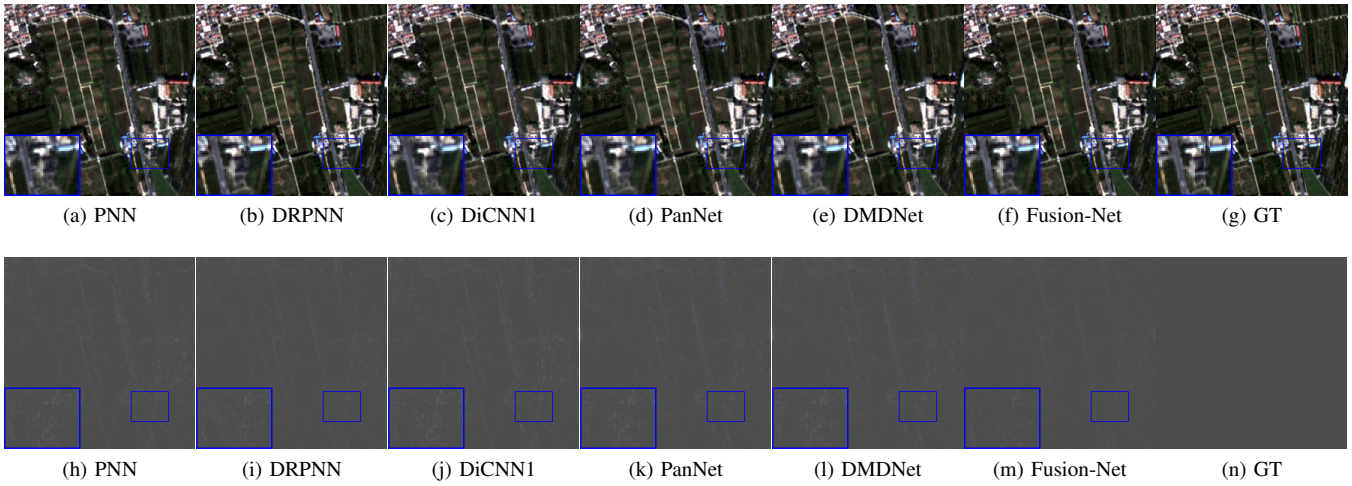
I. Discussions

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 and the giga floating-point operations per second (GFLOPs).

Detail Images: Unlike the previously shown absolute error maps, Fig. 16 displays the detail images in order to point out the differences among the compared methods. The detail 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.

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 the lower level than those of the other approaches, which demonstrates that the Fusion-Net gets the better training effectiveness.

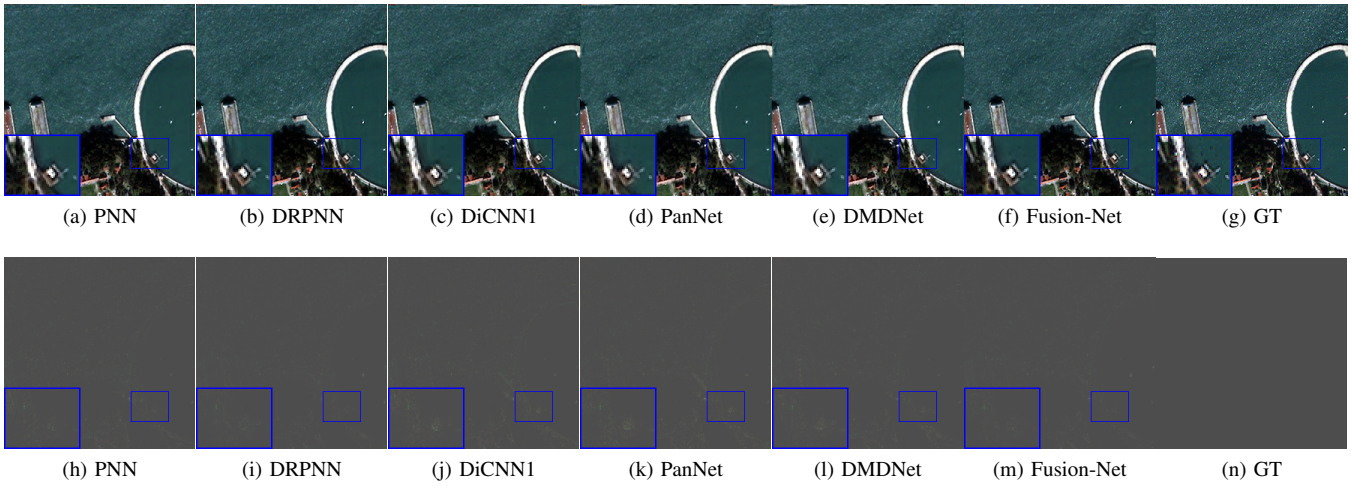
Network complexity: The proposed Fusion-Net is simpler than the PanNet. Comparing it with PanNet, Fusion-Net does not need to calculate the high-pass filtered version of the PAN image, thus reducing the training time with respect to PanNet. The architecture of DMDNet is similar as that of the PanNet, but DMDNet has a structure of grouped dilated convolution 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 3-layer network meaning that is not easy to extract sufficient image features. The architecture of PNN is a simple 3-layer network without any skip connection, thus it is also not easy to extract enough image features from the simple network. Additionally, the architecture of the DRPNN



(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. 14: Visual comparisons in natural colors of the most representative 6 approaches on the Guangzhou dataset (sensor: GF-2). First row: visual results; Second row: absolute error maps.



(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 6 approaches on the Indianapolis dataset (sensor: QB). First row: visual results; Second row: absolute error maps.

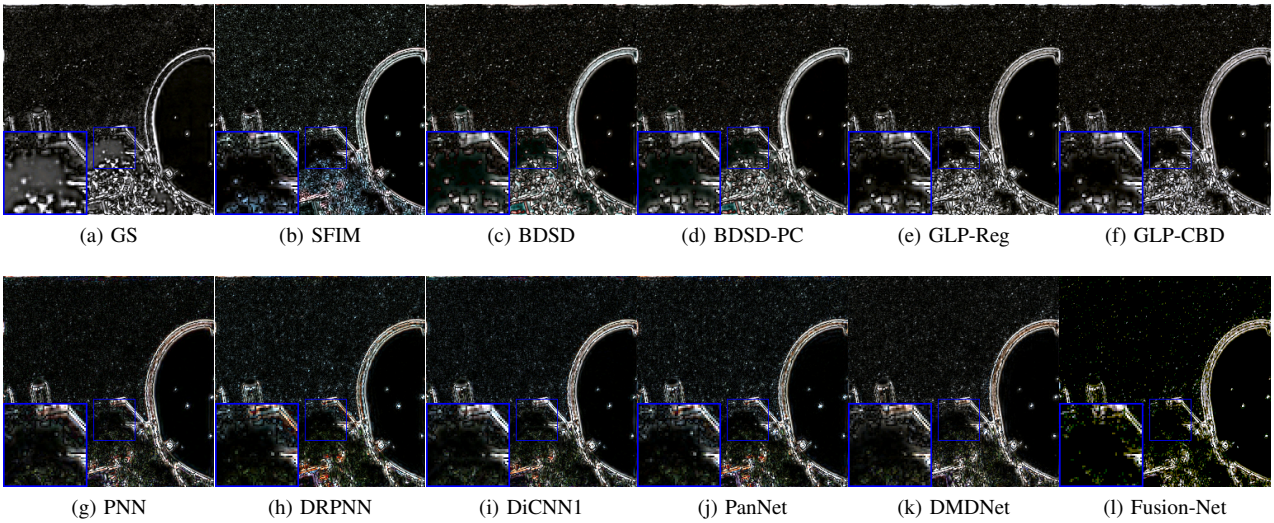
contains a skip connection and 11 layers, thus having a better feature extraction ability.

Testing time: Tab. III reports the testing time of all the compared methods on two WorldView-3 data (*i.e.*, Rio dataset and Tripoli dataset, both with size $256 \times 256 \times 8$). Classical CS and MRA methods generally reach shorter testing time than that of the CNN methods. Furthermore, it is worth to be noted that CNN times are calculated on a 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 propose networks can be considered acceptable on these data.

Training time: The training time of all the CNNs are reported using the same training dataset. The maximum iteration for each method is the optimal one used in the training phase. In Tab. VII, the proposed Fusion-Net yields the shortest training time mainly due to the less iterations when reaching convergence.

The number of parameters and GFLOPs: The number of parameters (NoPs) and the GFLOPs of all the compared CNNs are reported in Tab. VIII. From Tab. VIII, it is clear that the DiCNN gets the best performance on the NoPs and the GFLOPs, thanks 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.

Optimal iteration number for Fusion-Net: We want to investigate on the optimal value of the iteration number for the proposed Fusion-Net. In order to select it, we consider an exemplary reduced resolution dataset as Rio dataset. We calculated the performance metrics (the average of 5 runs) as in Fig. 18 taking the number of iterations that shows the



(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. 16: Detail images of the different compared methods on a sample belonging to the Indianapolis dataset (sensor: QB).

TABLE III: Quantitative results for Rio dataset and Tripoli dataset (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

best overall quality. Thus, we refer to the value that gets the maximum Q8 index (around 140000 iterations in Fig. 18), thanks to the fact that the Q8 can be considered an overall quality index. However, all the reduced resolution performance metrics are often in agreement among each other, see again Fig. 18.

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

V. Conclusions

We investigated in this paper on new architectures of convolutional neural networks for pansharpening. In particular, we focused our attention on deep convolutional neural networks 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 machine learning 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) compering it with the other CNN-based methods.

TABLE V: Quantitative results on Stockholm dataset (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

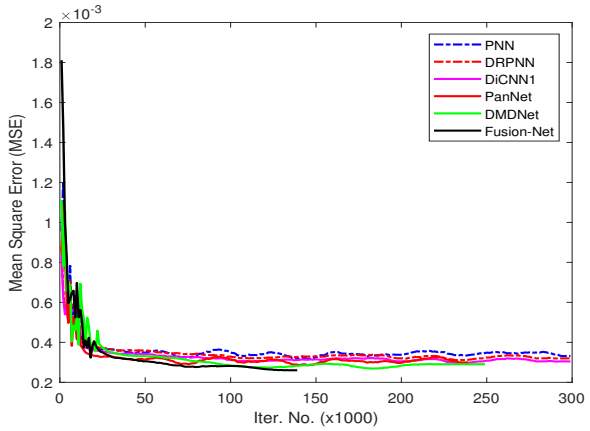


Fig. 17: Convergence curves for all the compared CNN methods on WorldView-3 training dataset. Note that we trained the PNN method with 1.12×10^6 iterations, but here we only show MSEs of the first 3×10^5 iterations for better display.

VI. ACKNOWLEDGMENT

L. -J. Deng thanks to NSFC (61702083, 61772003, 61876203) for partial support.

REFERENCES

[1] J. Yang, X. Fu, Y. Hu, Y. Huang, X. Ding, and J. Paisley, "PanNet: A deep network architecture for pan-sharpening," *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[2] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 4, pp. 1188–1204, 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 Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 3012–3021, 2007.

[4] M. Dalla Mura, S. Prasad, F. Pacifici, P. Gamba, and J. Chanussot, "Challenges and opportunities of multimodality and data fusion in remote sensing," *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, pp. 106–110, 2014.

[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 Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 1301–1312, 2008.

[6] G. Vivone, L. Alparone, J. Chanussot, M. Dalla Mura, A. Garzelli, G. Licciardi, R. Restaino, and L. Wald, "A critical comparison among pansharpening algorithms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 2565–2586, 2015.

[7] C. Souza, L. Firestone, L. Silva, and D. Roberts, "Mapping forest degradation in the Eastern Amazon from SPOT 4 through spectral mixture models," *Remote Sensing of Environment*, vol. 87, pp. 494–506, 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 Sensing of Environment*, vol. 199, pp. 241–255, 2017.

[9] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, pp. 228–236, 2008.

[10] G. Vivone, "Robust band-dependent spatial-detail approaches for panchromatic sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6421–6433, 2019.

[11] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution based satellite image fusion by using partial replacement," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, pp. 295–309, 2011.

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

[13] G. J. Liu, "Smoothing filter based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, pp. 3461–3472, 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 Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 2376–2385, 2005.

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

[16] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 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 Transactions on Geoscience and Remote Sensing*, vol. 40, pp. 2300–2312, 2002.

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

[19] G. Vivone, S. Marano, and J. Chanussot, "Pansharpening: Context-based generalized laplacian pyramids by robust regression," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–16, 2020.

[20] G. Vivone, R. Restaino, and J. Chanussot, "Full scale regression-based injection coefficients for panchromatic sharpening," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3418–3431, 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 Transactions on Image Processing*, vol. 23, pp. 4160–4174, 2014.

[22] Y. Jiang, X. Ding, D. Zeng, Y. Huang, and J. Paisley, "Pan-sharpening with a Hyper-Laplacian penalty," *International Conference on Computer Vision (ICCV)*, pp. 540–548, 2015.

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

[24] M. Moller, T. Wittman, and A. L. Bertozzi, "A variational approach to hyperspectral image fusion," *The International Society for Optical Engineering (SPIE)*, vol. 7334, pp. 73341E.1¨C73341E.10, 2009.

[25] F. Fang, F. Li, C. Shen, and G. Zhang, "A variational approach for pan-sharpening," *IEEE Transactions on Image Processing*, vol. 22, pp. 2822–2834, 2013.

[26] J. Duran, A. Buades, B. Coll, and C. Sbert, "A nonlocal variational model for pansharpening image fusion," *SIAM Journal on Imaging Sciences*, vol. 7, pp. 761–796, 2014.

[27] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 318–322, 2014.

[28] H. A. Aly and Sharma, "A regularized model-based optimization framework for pan-sharpening," *IEEE Transactions on Image Processing*, vol. 23, pp. 2596–2608, 2014.

TABLE VI: Quantitative assessment of the compared networks for the GF-2 testing dataset (81 samples) and the QB testing dataset (48 samples). Best results in boldface.

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

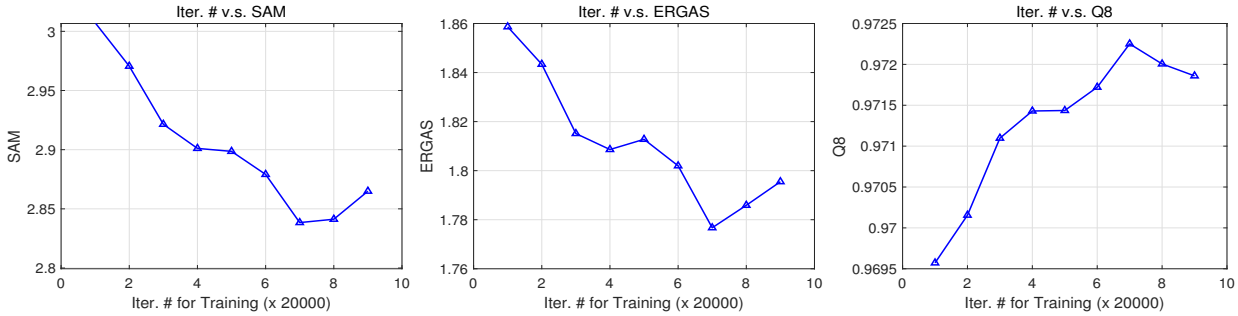


Fig. 18: Iteration number against quality metrics by averaging 5 runs on Rio dataset for the proposed Fusion-Net.

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 number of parameters (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**

[29] C. Chen, Y. Li, W. Liu, and J. Huang, "SIRF: Simultaneous satellite image registration and fusion in a unified framework," *IEEE Transactions on Image Processing*, vol. 24, pp. 4213–4224, 2015.

[30] G. Vivone, M. Simoes, M. Dalla Mura, R. Restaino, J. M. Bioucas-Dias, G. A. Licciardi, and J. Chanussot, "Pansharpening based on semiblind deconvolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 4, pp. 1997–2010, 2015.

[31] Q. Wei, N. Dobigeon, J. Y. Tourneret, J. M. Bioucas-Dias, and S. Godsill, "R-FUSE: Robust fast fusion of multi-band images based on solving a Sylvester equation," *IEEE Signal Processing Letters*, vol. 23, pp. 1632–1636, 2016.

[32] L. Deng, G. Vivone, W. Guo, M. Dalla Mura, and J. Chanussot, "A variational pansharpening approach based on reproducible kernel Hilbert space and heaviside function," *IEEE International Conference on Image Processing (ICIP)*, pp. 535–539, 2017.

[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 Sensing*, vol. 10, no. 4, pp. 622, 2018.

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

[35] G. Vivone, P. Addesso, R. Restaino, M. Dalla Mura, and J. Chanussot, "Pansharpening based on deconvolution for multi-band filter estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 1, pp. 540–553, 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," *Information Fusion*, vol. 52, pp. 76–89, 2019.

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

[38] X. X. Zhu and R. Bamler, "A sparse image fusion algorithm with application to pan-sharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, no. 5, pp. 2827–2836, 2013.

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

[40] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 1037–1041, 2015.

[41] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sensing*, vol. 8, pp. 594, 2016.

[42] Y. Rao, L. He, and J. Zhu, "A residual convolutional neural network for pan-shaprening," *International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, pp. 1–4, 2017.

[43] N. Li, N. Huang, and L. Xiao, "PAN-Sharpening via residual deep learning," *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5133–5136, 2017.

[44] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network,"

IEEE Geoscience and Remote Sensing Letters, vol. 14, no. 10, pp. 1795–1799, 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 Transactions on Cybernetics*, vol. 49, no. 7, pp. 2406–2419, 2018.

[46] Y. Xu, L. Zhang, B. Du, and F. Zhang, "Spectral-spatial unified networks for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 10, pp. 5893–5909, 2018.

[47] Z. Shao, Z. Lu, M. Ran, L. Fang, J. Zhou, and Y. Zhang, "Residual encoder-decoder conditional generative adversarial network for pan-sharpening," *IEEE Geoscience and Remote Sensing Letters*, pp. 1–5, 2019.

[48] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive cnn-based pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 9, pp. 5443–5457, 2018.

[49] S. Eghbalian and H. Ghassemian, "Multi spectral image fusion by deep convolutional neural network and new spectral loss function," *International Journal of Remote Sensing*, vol. 39, no. 12, pp. 3983–4002, 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 3, pp. 978–989, 2018.

[52] L. Liu, J. Wang, E. Zhang, B. Li, X. Zhu, Y. Zhang, and J. Peng, "Shallow-deep convolutional network and spectral-discrimination-based detail injection for multispectral imagery pan-sharpening," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 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," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[54] W. Xie, Y. Cui, Y. Li, J. Lei, Q. Du, and J. Li, "HPGAN: Hyperspectral pansharpening using 3-d generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–15, 2020.

[55] C. Dong, C. C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 295–307, 2016.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[57] L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images," *Photogrammetric Engineering and Remote Sensing*, vol. 63, pp. 691–699, 1997.

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

[59] G. Vivone, R. Restaino, M. Dalla Mura, G. Licciardi, and J. Chanussot, "Contrast and error-based fusion schemes for multispectral image pansharpening," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, pp. 930–934, 2014.

[60] Y. Fu, W. Wang, Y. Huang, X. Ding, and J. Paisley, "Deep multi-scale detail networks for multi-band spectral image sharpening," *IEEE Transactions on Neural Networks and Learning Systems*, 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," *JPL Airborne Geoscience Workshop; AVIRIS Workshop: Pasadena, CA, USA*, pp. 147–149, 1992.

[62] L. Wald, "Data fusion: definitions and architectures: Fusion of images of different spatial resolutions," *Presses 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," *International Journal of Remote Sensing*, vol. 19, pp. 743–757, 1998.

[64] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of multi-/hyper-spectral images," *IEEE Geoscience and Remote Sensing Letters*, vol. 6, no. 4, pp. 662–665, 2009.