applied sciences

MDPI

Article

Coupled Convolutional Neural Network-Based Detail Injection Method for Hyperspectral and Multispectral Image Fusion

Xiaochen Lu [1], Dezheng Yang [1], Fengde Jia [1,*] and Yifeng Zhao [2]

[1] School of Information Science and Technology, Donghua University, Shanghai 201620, China; lxchen09@dhu.edu.cn (X.L.); dezheng.yang@mail.dhu.edu.cn (D.Y.)
[2] Shanghai Radio Equipment Research Institute, Shanghai 201109, China; zhaoyf_hit@163.com
* Correspondence: fdjia@dhu.edu.cn; Tel.: +86-0136-9944-1635

Featured Application: This work proposes a CNN-based hyperspectral and multispectral image fusion method, which aims at improving the spatial resolution of hyperspectral image, thereby contributing to the accurate identification and classification of land-covers.

Abstract: In this paper, a detail-injection method based on a coupled convolutional neural network (CNN) is proposed for hyperspectral (HS) and multispectral (MS) image fusion with the goal of enhancing the spatial resolution of HS images. Owing to the excellent performance in spectral fidelity of the detail-injection model and the image spatial–spectral feature exploration ability of CNN, the proposed method utilizes a couple of CNN networks as the feature extraction method and learns details from the HS and MS images individually. By appending an additional convolutional layer, both the extracted features of two images are concatenated to predict the missing details of the anticipated HS image. Experiments on simulated and real HS and MS data show that compared with some state-of-the-art HS and MS image fusion methods, our proposed method achieves better fusion results, provides excellent spectrum preservation ability, and is easy to implement.

Keywords: convolutional neural network; hyper-sharpening; hyperspectral; image fusion; multispectral

Citation: Lu, X.; Yang, D.; Jia, F.; Zhao, Y. Coupled Convolutional Neural Network-Based Detail Injection Method for Hyperspectral and Multispectral Image Fusion. *Appl. Sci.* **2021**, *11*, 288. https://doi.org/10.3390/app11010288

Received: 9 December 2020
Accepted: 28 December 2020
Published: 30 December 2020

1. Introduction

Hyperspectral (HS) imagery presents plentiful spectral details and allows for accurate analyses of terrestrial features due to its high spectral resolution. However, owing to the constraint of sensors, HS images usually have low spatial resolution, which limits its applications in some circumstances. During recent decades, the increased application requirement has encouraged the demand for improved potential spatial resolution of HS images. By employing image fusion techniques, HS and high-resolution multispectral (MS) images thus have a possibility to produce enhanced HS data that would contribute to the accurate identification and classification of land-covers observed at a finer ground resolution.

For this purpose, hyper-sharpening [1], i.e., fusion of HS and MS images, has attracted considerable concern over the past decade. In the early years, MS image pan-sharpening techniques were used to cope with some simple hyper-sharpening tasks. To tackle the problem of high spectral fidelity demand, the generalization of pan-sharpening algorithms has shown substantial successes in the HS image area [2–4]. In [5], the maximum a posteriori (MAP) estimation is presented for fusing HS data with an auxiliary high-resolution image, which can be an MS or panchromatic (PAN) image. To achieve better noise tolerance and fast implementation time, Refs. [6,7] propose to incorporate wavelet transform and principal component analysis with the MAP estimation method. In addition, a fast fusion algorithm for multi-band images is clarified in [8], applying a Sylvester equation. Similarly, Ref. [9] adopts a fast-multi-band image fusion algorithm with robustness. In [10], by formulating data fusion as a convex optimization problem, a "HySure" algorithm is proposed to use a form of vector total variation-based regularization method to cope with

HS sharpening and super-resolution problems. Sparse representation [11,12] is also used to redefine the fusion problem for HS and MS images and shows excellent enhancement, especially in spatial aspects.

The matrix factorization-based hyper-sharpening techniques draw worldwide interest, due to the report of the coupled nonnegative matrix factorization (CNMF) algorithm [13], which alternately unmixes HS and MS images into endmember and abundance matrices based on a linear spectral mixture model. Inspired by CNMF, the dictionary-pair learning method (DPLM) [14] and many other matrix factorization methods are proposed successively. For instance, two approaches are proposed in [15], i.e., gradient-based joint-criterion nonnegative matrix factorization (JNMF) and multiplicative JNMF method. Ref. [16] proposes the spectral modulation hyper-sharpening methods, aiming at minimizing the spectral distortion of real MS and HS images acquired by different sensors or platforms. Considering the spectral variability of remote sensing scenes acquired at different times, [17] combines the unmixing-based formulation with an explicit parametric model to implement HS and MS image fusion. Compared with those conventional methods, matrix factorization-based methods show excellent spectral fidelity and capability of anti-noise. In addition, tensor-based approaches have also been explored to further improve the performance over matrix-based techniques, since they generally regard HS image as a 3D tensor rather than multiple 2D matrices and jointly exploit the spatial-spectral structure information. Refs. [18–22] have made comprehensive analyses and comparisons about those classical hyper-sharpening algorithms.

In recent years, with the development and increased application of deep learning techniques, convolutional neural networks (CNNs) [23] were widely applied to those image-related tasks owing to their outstanding adaptability and practicability in exploring and extracting local spatial structure characteristics. Accordingly, several CNN models have been proposed to deal with tasks related to image fusion or super-resolution: Refs. [24–27] implement the fusion work by utilizing a 3-D convolution neural network, with the dimension of HS image reduced beforehand. In [28], a deep HS sharpening method, abbreviated to DHSIS, is presented to learn the image priors via deep CNN-based residual learning. Ref. [29] proposes a two-branches CNN fusion method (abbreviated to TCNNF for convenience in this paper), which explores the features from the spectrum of each pixel in low-resolution HS images and its corresponding spatial neighborhood in MS images by 1-D and 2-D CNN branches, respectively. The extracted features are subsequently concatenated and fed to fully connected (FC) layers. In this way, the spatial and spectral information of HS and MS images could be fully fused. In [30], a pyramid fully convolutional network made up of an encoder sub-network and a pyramid fusion sub-network are proposed to refine the spatial information of the multispectral image in a global-to-local manner. Ref. [31] proposes an HS and MS image fusion method (called as CNN-Fus), which is based on subspace representation and CNN denoiser. Ref. [32] proposes a novel variational probabilistic autoencoder framework implemented by CNN in order to fuse the spatial and spectral information contained in the low-resolution (LR) HS and high-resolution (HR) MS images. This method is called FusionNet. To encode spatial and spectral distortion, Ref. [33] proposes a complex multi-scale fusion neural network, termed HAM-MFN, which designs two branches to extract features of low-resolution HS and MS images, respectively, and then fuse them at different scales. In [34], a quadratic optimization network with matrix decomposition is constructed, and the fusion problem is substituted by the optimization problem for spectrum and space with a customized loss function. Finally, in [35], authors propose a new framework, called recurrent attention fusion network (abbreviated to RAFnet) to obtain high-resolution HS images in an unsupervised manner.

In pan-sharpening area, the detail-injection-based methods usually show high color preservation ability and also have considerable potential in hyper-sharpening tasks [36] naturally. Meanwhile, deep learning, especially the convolutional neural network, offers impressive performance of non-linear and local-structure feature extraction in the computer

vision field. Therefore, inspired by the detail-injection pan-sharpening methods proposed in [37], in this paper, we propose a coupled CNN-based detail injection method (abbreviated to CpCNN) for HS and MS image fusion. The proposed method employs a couple of convolutional neural networks for high-frequency detail feature extraction and prediction of HS and MS images, respectively. Compared with conventional detail injection methods, a CNN-based model learns the spatial structure information from the pending images in an automatic and supervised fashion, thereby circumventing the intermediate process of separately estimating the details and injection gains and reducing the model uncertainty. The main contributions of this paper are as follows: (1) it is the first effort that incorporates a CNN network with a detail-injection model for the hyper-sharpening task, which would contribute to a substantial improvement in spectral fidelity; (2) in contrast with pan-sharpening works, we use a couple of fully 2-D CNNs to separately exploit the collaborative spatial and spectral characteristics of HS and MS images and jointly predict the missing details of high-resolution HS image; (3) by adopting different sizes of image patches and numbers of convolutional layers, the structure features of HS and MS images within different scales can be automatically explored. Similar to the TCNNF method, the output of our network involves the details of each individual pixel, which avoids the potential errors that might be caused by overlaps of multiple patches.

The remainder of this paper is organized as follows. Section 2 introduces the study datasets and elaborates the presented CNN-based detail injection method. Section 3 describes the experiments and results. Finally, the conclusion is drawn in Section 4.

2. Materials and Methods

2.1. Detail Injection Sharpening Framework

Given the observable low-spatial-resolution HS image $\mathbf{X} \in \mathbb{R}^{m \times n \times \Lambda}$ and the high-spatial-resolution MS image $\mathbf{Y} \in \mathbb{R}^{M \times N \times \lambda}$, where m, n, M, and N are the rows and columns of images respectively, and Λ and λ are the number of bands, (generally, $M > m$, $N > n$, and $\Lambda > \lambda$), $s = M/m$ represents the spatial resolution ratio, and the goal of HS and MS image fusion is to estimate the unobservable HS image $\mathbf{Z} \in \mathbb{R}^{M \times N \times \Lambda}$ with high spatial and spectral resolutions. Generally, the LR HS image can be spatially upscaled to the size of $M \times N \times \Lambda$ by bilinear interpolation method [2,38], which can be denoted as $\boldsymbol{X}^{\mathrm{up}} \in \mathbb{R}^{M \times N \times \Lambda}$. Therefore, the difference between the LR and HR HS images is the lack of HR spatial details, which makes the LR HS image have blurred edges and textures. The detail injection model aims to estimate the missing details of the HS image, and to inject into the LR HS image directly to obtain a new HS image that would be observed at the same resolution with the HR MS image. Technically, a detail-injection model can be described as follows:

$$\widetilde{\boldsymbol{Z}}_k = \boldsymbol{X}_k^{\mathrm{up}} + \boldsymbol{D}_k, k = 1, 2, \ldots, \Lambda \tag{1}$$

where $\widetilde{\boldsymbol{Z}}$ is the spatially enhanced HS image, $\boldsymbol{D}$ denotes the residual image, i.e., the missing detail component, which can be generally computed from the corresponding MS image bands, e.g.,

$$\boldsymbol{D}_k = \sum_{i=1}^{\lambda} \alpha_{ik} \cdot \left(\boldsymbol{Y}_i - \boldsymbol{Y}_i^{\mathrm{low}} \right), k = 1, 2, \ldots, \Lambda \tag{2}$$

where $\boldsymbol{Y}^{\mathrm{low}}$ is the spatially degraded MS image that can be simulated by several approaches [39]. α_{ik} is the injection gain associated with the corresponding i-th MS band. Thus, we have

$$\widetilde{\boldsymbol{Z}}_k = \boldsymbol{X}_k^{\mathrm{up}} + \sum_{i=1}^{\lambda} \alpha_{ik} \cdot \left(\boldsymbol{Y}_i - \boldsymbol{Y}_i^{\mathrm{low}} \right), k = 1, 2, \ldots, \Lambda \tag{3}$$

Conventional detail injection methods usually employ a multi-resolution analysis (MRA) model to obtain the high-frequency detail component in a band-by-band manner, which highly depends on the selection and performance of the MRA model [14]. In particular, for hyper-sharpening tasks, due to the large difference in spectral resolution between

the HS and MS images, it is difficult to accurately predict the missing details by manually selecting the MRA model and computing the injection gains under various circumstances. By contrast, a convolutional neural network has the ability of learning and predicting the spatial structure features from the pending images in an automatic and supervised fashion. The details are driven from the context, which effectively reduces the model uncertainty and achieves higher image quality and adaptability. Hence, in this paper, we propose to utilize CNN to automatically learn the spatial details from HS and MS images themselves to promote the performance of image fusion.

2.2. Proposed Coupled CNN Fusion Approach

The proposed CNN-based fusion method can be summarized in Figure 1, in which two CNN networks with different numbers of convolutional layers are applied to HS and MS image patches, respectively. In practice, the HS image will be first upscaled to the size of the MS image by bilinear interpolation as mentioned above. Afterwards, both HS and MS image patches are selected by partitioning the two images through a pixel-by-pixel fashion with a stride 1×1 and fed into the two sub-networks, respectively. As a matter of fact, MS image has such a higher spatial resolution than HS images that they are substantially subject to different spatial scales. Therefore, the input of our network is composed of different spatial and spectral sizes of HS and MS image patches, which can be denoted by $p_H \times p_H \times \Lambda$ and $p_M \times p_M \times \lambda$, $p_H < p_M$ and $\Lambda > \lambda$ in general in order to fully exploit the structure information of HR image and spectral information of HS images.

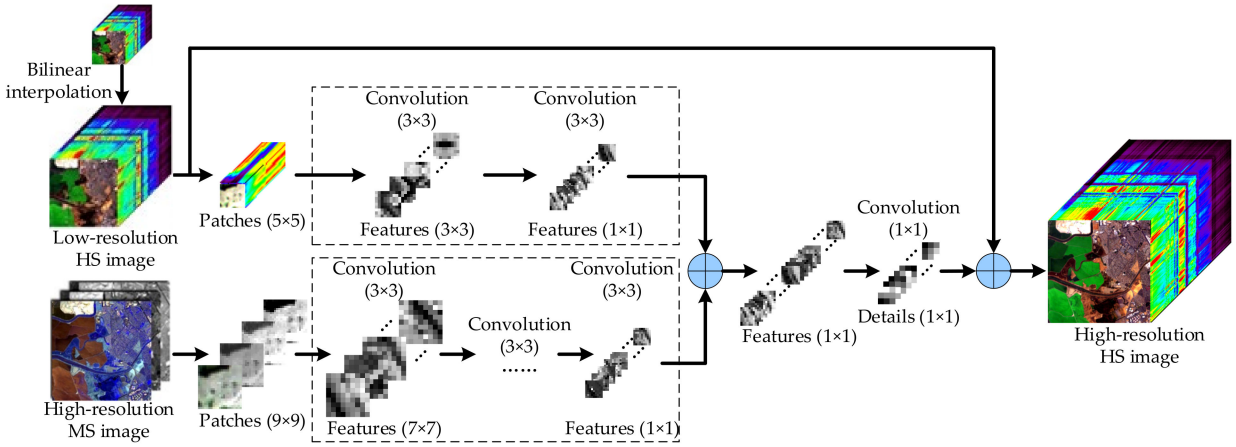

Figure 1. The main framework of the proposed method.

For the HS and MS sub-networks, different numbers of convolutional layers are applied on the image patches, respectively, which can be generally determined as $[(p_H\text{-}1)/2]$ and $[(p_M\text{-}1)/2]$, respectively, where $[\cdot]$ denotes the largest integer that is less than or equal to the given number. Each convolutional layer is followed by a batch normalization (BN) layer and an active function with a rectified linear unit (ReLU), individually. Therefore, the response of the l-th layer can be denoted as

$$\boldsymbol{A}^l = max\left(\boldsymbol{BN}\left(\boldsymbol{W}^l * \boldsymbol{A}^{l-1} + \boldsymbol{b}^l\right), 0\right) \tag{4}$$

where $\boldsymbol{W}^l$ and $\boldsymbol{b}^l$ are the weight and bias matrices of l-th layer; $\boldsymbol{A}^l$ denotes the output of l-th layer; $*$ denotes the convolution operation; BN$(\cdot)$ denotes the batch normalization operation, which is used to overcome the issue of internal covariate shift and accelerate the training of network. Consequently, both sub-networks yield a 1×1 size of output feature maps. Typically, in this paper, the patch sizes of HS and MS images are $5 \times 5 \times \Lambda$ and $9 \times 9 \times \lambda$, respectively, thus the numbers of convolutional layers are 2 and 4, respectively.

For both sub-networks, the convolutional filter sizes are all set to 3×3, which is commonly used by related works, and the convolutional filter numbers are set to 32.

Subsequently, the output feature maps of both HS and MS sub-networks are concatenated to form 64 feature maps and fed into an additional convolutional layer with Λ filters and 1×1 filter size to predict the residuals, resulting in $1 \times 1 \times \Lambda$ values for each pixel. The residuals, namely the predicted details, which are denoted by $\boldsymbol{D}_k$, $k = 1, 2, \ldots, \Lambda$, are finally added to the primary HS image to reconstruct the expected HR HS image.

It should be noted that the input patches of HS and MS sub-networks include the neighbor block of a pixel, and the output of the network is the residuals of the current pixel. In this way, the network will automatically utilize the collaborative spatial and spectral information of HS and MS images, respectively, and infer the details of the current pixel accurately.

In order to effectively train the network, both HS and MS images are decimated by the factor s ($s = M/m$). Suppose the decimated HS image is denoted by $\boldsymbol{X}^{\text{low}} \in \mathbb{R}^{(m/s) \times (n/s) \times \Lambda}$. As mentioned above, $\boldsymbol{X}^{\text{low}}$ will first be upscaled to the size of the decimated MS image by bilinear interpolation. For conciseness, it is also denoted by $\boldsymbol{X}^{\text{low}}$, namely $\boldsymbol{X}^{\text{low}} \in \mathbb{R}^{m \times n \times \Lambda}$. For each pixel, the details consist of $1 \times 1 \times \Lambda$ values, corresponding to the Λ spectral bands of the HS image. Thus, the expected total details of the HS image at lower scale consist of $m \times n \times \Lambda$ values and can be denoted by $\boldsymbol{D}^{\text{low}} \in \mathbb{R}^{m \times n \times \Lambda}$. Then, the injection model at lower scale can be transformed into

$$\boldsymbol{D}_k^{\text{low}} = \boldsymbol{X}_k - \boldsymbol{X}_k^{\text{low}}, k = 1, 2, \ldots, \Lambda \tag{5}$$

Therefore, the network is used to predict the expected spatial details of HS image at the lower scale; consequently, the loss function of the network can be formulated by,

$$J = E\left(\left\| \boldsymbol{D}^{\text{low}} - \widetilde{\boldsymbol{D}}^{\text{low}} \right\|_F^2 \right) = \frac{1}{mn} \sum_{i=1}^{mn} \left(\sum_{k=1}^{\Lambda} \left(d_{ik}^{\text{low}} - \tilde{d}_{ik}^{\text{low}} \right)^2 \right) \tag{6}$$

where the subscript i denotes the i-th patches (i.e., the i-th pixel); $\widetilde{\boldsymbol{D}}^{\text{low}}$ denotes the output details of the last layer of the network at the lower scale. The network can be trained by stochastic gradient descent with the backpropagation method. Once the network is trained, the primary HS image will be upscaled to the size of the primary MS image as aforementioned (i.e., $\boldsymbol{X}^{\text{up}}$), and thus, the network will be used to infer the high-level details to reconstruct the anticipated HR HS image according to (1).

3. Experimental Results and Discussion

In order to validate the effectiveness of our presented method, in this section, we report the experiments and results tested on four datasets.

3.1. Experimental Setup

The experimental datasets include three well-known individual HS images and a group of spatially co-registered real HS and MS images. For detailed information on these images, we recommend readers to refer to the following literature: [13,15,39].

The selected subsets of the first three HS images with 320×320, 400×400, and 400×240 pixels, respectively. To simulate the LR HS images, the HS images were spatially blurred by Gaussian low-pass filters with 3×3, 7×7, and 11×11 sizes, followed by downsampling operations with stride 2×2, 4×4, and 6×6, respectively, whereas the HR MS images were simulated by averaging the original HS bands with the wavelength covered by the corresponding MS bands of Landsat 5 TM sensor. In addition, Gaussian noise with peak-signal-to-noise of 40 dB was added to the MS image to simulate the different imaging conditions of HS and MS images. The last HS image includes 200×200 pixels and 159 bands after removing the noisy bands. The corresponding MS image acquired by the ASTER imager consisted of three bands and was blurred by applying a Gaussian filter

with a 3 × 3 size and down-sampled to the original resolution of the HS image. The LR HS image is then simulated by similar operations with the first three datasets.

In our experiments, several state-of-the-art hyper-sharpening methods, including CNMF [13], JNMF [15], DPLM [14], and the CNN-based algorithm TCNNF [29], were used to compare with our CpCNN method. We also adapted the detail-injection-based CNN pan-sharpening method [38] for our hyper-sharpening task (abbreviated to DiCNN) to further validate the superiority of our method. According to the related references [13,15], 40 end-members were extracted by vertex component analysis for CNMF and JNMF. The dictionary dimensionality of DPLM is also 40, and the optimal sparseness degrees were set to 0.7 for the second dataset and 0.6 for others. For TCNNF and DiCNN, the parameters were set mostly in accordance with the authors' suggestion, including the number of convolutional layers and filters, learning rate, momentum, etc. For fairness, in our CpCNN method, apart from the filter and patch sizes, parameters were determined basically according to the TCNNF and DiCNN methods, e.g., the learning rate was fixed as 0.0001, the momentum was set to 0.9, the batch size was 128, and the number of training epochs was 200. The filter number and size of each convolutional layer were set to 32 and 3 × 3, respectively, which are illustrated in Section 2. In particular, the last convolutional layer including Λ filters had a 1 × 1 size, corresponding to the details of Λ bands for a single pixel. All the weights of convolutional filters were initialized by Gaussian random distributions with a zero-mean and standard deviation of 0.01.

To achieve the best performance, the MS image patches of TCNNF and the HS/MS image patches of DiCNN were composed of 21 × 21 pixels, which was determined empirically according to our experiments. For our CpCNN method, the patch sizes of HS and MS images were set to 5 × 5 and 9 × 9, respectively, in the following sub-section. Specifically, for these CNN-based methods, including our CpCNN method, the samples (namely image patches) were generated by selecting each pixel with its neighbors of the HS and MS images, respectively. As mentioned above, the degraded HS image is first interpolated to the size of the MS image. Therefore, for a given degraded HS or MS image with n pixels, a total of n training samples, namely n pixels, are used to train the network. The testing set, thus, has N samples, corresponding to N pixels of the HR HS and MS images.

The experiments were conducted on Windows 10 operating system based on Anaconda toolkit with Python 3.7. An NVIDIA GeForce RTX 2060 GPU card was used to train the CNN networks. Several widely used indices, including the spectral angle mapper (SAM), relative dimensionless global error in synthesis (ERGAS), and universal image quality index (UIQI), were adopted to evaluate the fusion performance in order to give a comprehensive assessment. For the sake of avoiding stochastic errors caused by random initialization of network weights, 5 independent runs were conducted for CNN network training and testing, and the average values were calculated and displayed in the following sub-sections.

3.2. Results and Analyses

In this sub-section, we report the experimental results conducted on the four datasets. The experiments were conducted under different resolution ratios, i.e., the HS image was down-sampled by the facts of 2, 4, and 6 (i.e., s = 2, 4, and 6), respectively, as Section 3.1 mentioned. The numerical results are listed in Tables 1–3, respectively, where the best values are highlighted in bold. For CNN-based approaches, the training and testing times are listed separately in Tables 1–3. By observing the SAM, ERGAS, and UIQI values, we can see that among the conventional matrix factorization-based methods, JNMF and CNMF provide similar performance for these datasets on the whole, whereas DPLM has an impressive ability to alleviate the spectral distortion that might be caused by different imaging conditions, since it had outstanding performance on the San Francisco dataset. The CNN-based algorithms, namely the last three methods, obviously surpass the conventional methods in totally minimizing the fusion errors, as in most cases, they achieved far more desirable results. For example, the SAM and ERGAS values of San Francisco dataset under

a resolution ratio of 2 are much lower than the former three methods, whereas the UIQI values are much higher. The presented CpCNN approach exhibits the best spectral fidelity performance, since in most cases, it achieves lower fusion errors in both angle (spectrum shape) and amplitude aspects and higher image quality.

By comparing Tables 1–3, it can be clearly seen that with the decrement of HS image resolution (i.e., s varies from 2 to 6), the qualities of fused images undoubtedly degenerate. Nevertheless, the CNN-based approaches remain remarkable in performance, especially for the detail-injection methods. It should also be recognized that, in fact, the resolution decrement also results in a decrement of training samples, since the HS and MS images in the degraded scales contain substantially fewer pixels, which certainly affects the training of networks. In summary, these tables indicate that the proposed CpCNN is significantly preferable to the other approaches in spectral characteristic preservation aspect. Additionally, the running times in the tables suggest that JNMF seems to be the most time-saving method in our experiments. The CNN-based approaches are generally composed of training and testing stages. We can see that the training stage of the network always takes a considerably long time, which highly depends on the architecture of the network and the number of samples. Once the network is trained, we will not spend much time inferring the high-resolution maps. In addition, thanks to the small patch and filter sizes, the training time of our network is far less than TCNNF and DiCNN, although it is still time-consuming compared with the conventional methods.

Table 1. Numerical evaluation of hyper-sharpening results under resolution ratio of 2 (s = 2).

Method	Dataset	SAM	ERGAS	UIQI	Time (s)	Dataset	SAM	ERGAS	UIQI	Time (s)
CNMF	University of Pavia	4.12	5.70	0.9807	39.3	Pavia City center	10.11	7.51	0.9860	63.9
JNMF		4.59	5.65	0.9774	**14.2**		9.10	7.49	0.9858	**21.3**
DPLM		4.00	5.31	0.9826	49.9		10.72	8.64	0.9815	153.4
TCNNF		2.76	3.90	0.9911	3079.0/26.5		5.80	5.47	0.9925	4797.9/40.8
DiCNN		3.06	4.25	0.9902	10,553.1/105.8		5.60	5.67	0.9917	16,545.7/169.2
CpCNN		**2.62**	**3.47**	**0.9927**	1449.3/**8.5**		**5.32**	**4.69**	**0.9943**	2273.7/**13.1**
CNMF	Washington DC Mall	2.59	4.74	0.9777	43.4	San Francisco	8.89	16.31	0.8157	15.6
JNMF		3.21	6.16	0.9724	**15.2**		8.96	17.71	0.8155	5.2
DPLM		2.83	7.74	0.9703	56.9		4.33	10.78	0.9305	23.7
TCNNF		1.44	4.94	0.9904	3953.1/35.3		2.87	6.53	0.9747	1434.4/13.5
DiCNN		1.33	2.90	**0.9958**	13,573.1/149.7		2.45	5.56	0.9806	5140.8/53.2
CpCNN		**1.14**	**2.81**	**0.9958**	1596.3/**11.5**		**2.23**	**5.52**	**0.9819**	603.7/**4.6**

Table 2. Numerical evaluation of hyper-sharpening results under resolution ratio of 4 (s = 4).

Method	Dataset	SAM	ERGAS	UIQI	Time (s)	Dataset	SAM	ERGAS	UIQI	Time (s)
CNMF	University of Pavia	4.31	3.55	0.9709	38.3	Pavia City center	9.89	4.49	0.9797	61.1
JNMF		4.73	2.93	0.9756	**16.6**		9.13	3.82	0.9848	**26.7**
DPLM		4.23	2.94	0.9782	46.0		11.81	4.75	0.9780	164.8
TCNNF		3.75	2.73	0.9823	711.5/24.3		**6.75**	3.66	0.9865	1103.5/38.9
DiCNN		3.41	2.75	0.9843	2572.0/107.6		8.22	3.34	0.9891	4237.1/170.6
CpCNN		**3.34**	**2.26**	**0.9889**	382.2/**8.7**		6.81	**2.92**	**0.9915**	593.3/**13.3**
CNMF	Washington DC Mall	2.80	2.66	0.9708	41.0	San Francisco	8.88	8.19	0.8046	15.9
JNMF		3.57	3.30	0.9681	**17.2**		8.90	9.00	0.8030	**6.5**
DPLM		3.27	4.04	0.9666	52.8		4.81	5.74	0.9184	30.2
TCNNF		2.14	2.93	0.9862	917.1/33.8		4.29	5.28	0.9245	341.6/12.7
DiCNN		1.72	2.00	0.9922	3399.6/145.6		4.17	5.48	0.9415	1220.2/49.8
CpCNN		**1.62**	**1.97**	**0.9931**	444.6/**11.3**		**3.84**	**4.56**	**0.9485**	161.2/**4.6**

Figures 2–5 show the false-color images of fusion results under a resolution ratio of 4 (i.e., s = 4). For the first dataset, the numerical evaluation suggests that all of the fused images have high similarity with the reference, e.g., the largest SAM and ERGAS among different approaches are 4.73 and 3.55, respectively, whereas the worst UIQI is 0.9709; thus, we can see that the colors of these images in Figure 2 are close to the reference image

in [38]. Some minor differences can be observed on the red roof in the lower right corner, where CNMF has slight color distortion compared with the others. The more apparent discrepancies can be observed on the lake of Figure 3, where we can see that the last three images have similar colors with the reference image, corresponding to the higher UIQIs (0.9865, 0.9891, and 0.9915, respectively) and lower SAMs (6.75, 8.22, and 6.81) as well as the ERGASs (3.66, 3.34, and 2.92) values than the conventional methods.

Likewise, in Figures 4 and 5, CNN-based images usually present approximate spectral characteristics. Similarly, the SAMs achieve 2.14, 1.72, and 1.62, and the UIQIs are higher than 0.98, for the Washington DC Mall dataset, which performs much better than the former ones. Since the last dataset contains HS and MS images acquired by different platforms, the fused images exhibit large color distortions. Nonetheless, the numerical results still suggest that the proposed method has the best image quality, i.e., the SAM, ERGAS, and UIQI are 3.84, 4.56, and 0.9485, respectively, which are obviously superior to the others.

In order to give an intuitive comparison of spatial detail aspects, Figures 6 and 7 show some local enlargements of the fused images. Of course, compared with the simulated LR HS image in Figures 6d and 7d, the fused images present abundant spatial details and clearer texture characteristics. From Figure 6, it can be seen that the conventional methods always suffer from apparent noise, which is caused by the simulated noisy MS image. This can also be inferred from Table 2; e.g., the SAMs and ERGASs of CNMF and DPLM are 9.89, 11.81, 4.49, and 4.75, respectively. By contrast, the CNN-based methods can effectively overcome this issue. From Figure 7, we can see that the detail-injection CNN methods have relatively better overall appearances, corresponding to the lower SAMs (4.17 and 3.84) and higher UIQIs (0.9415 and 0.9485), whereas the TCNNF method presents obscure edges and textures compared with the other methods.

In sum, the CNN-based hyper-sharpening methods are much preferable to the conventional matrix factorization-based methods, and our proposed method indeed surpasses the other CNN-based methods in both quantity and visual aspects.

Table 3. Numerical evaluation of hyper-sharpening results under resolution ratio of 6 (s = 6).

Method	Dataset	SAM	ERGAS	UIQI	Time (s)	Dataset	SAM	ERGAS	UIQI	Time (s)
CNMF	University of Pavia	4.40	2.98	0.9554	25.10	Pavia City center	10.06	3.74	0.9680	39.53
JNMF		4.73	**1.97**	0.9751	**11.05**		9.62	2.95	0.9799	**17.21**
DPLM		5.19	4.40	0.9039	46.33		12.19	3.48	0.9735	180.01
TCNNF		4.62	3.09	0.9544	358.6/27.1		8.69	3.04	0.9798	556.5/40.6
DiCNN		**4.37**	2.01	**0.9800**	1205.3/106.0		9.45	2.56	0.9856	2110.6/168.0
CpCNN		4.55	2.46	0.9732	164.4/**8.5**		**7.50**	**2.45**	**0.9876**	255.9/**12.9**
CNMF	Washington DC Mall	4.87	2.51	0.9503	24.49	San Francisco	8.44	5.67	0.7811	9.93
JNMF		3.61	2.17	0.9708	**10.75**		8.85	5.89	0.8010	4.34
DPLM		3.49	2.81	0.9603	56.37		5.43	4.27	0.8959	22.75
TCNNF		2.43	2.79	0.9717	443.9/35.5		5.53	4.60	0.8823	179.6/13.9
DiCNN		3.14	2.46	0.9786	1538.4/148.4		8.01	5.06	0.8928	612.7/54.2
CpCNN		**1.91**	**1.71**	**0.9889**	181.9/**11.5**		**4.43**	**3.88**	**0.9171**	**72.9/4.5**

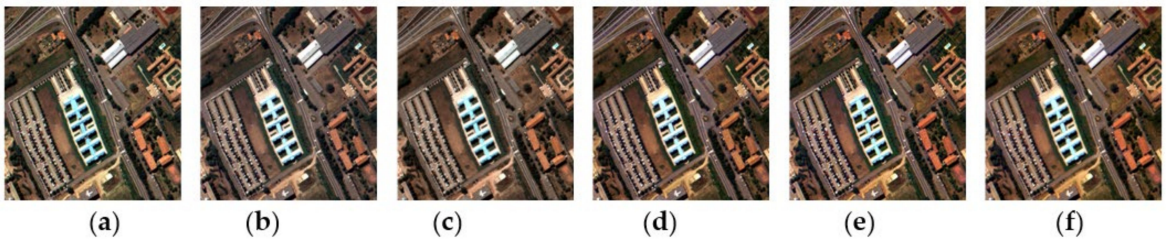

Figure 2. Fused images of University of Pavia. (**a**) coupled nonnegative matrix factorization (CNMF); (**b**) JNMF; (**c**) dictionary-pair learning method (DPLM); (**d**) two-branches CNN fusion method; (**e**) detail-injection-based CNN fusion method (DiCNN); (**f**) coupled CNN-based fusion method (CpCNN).

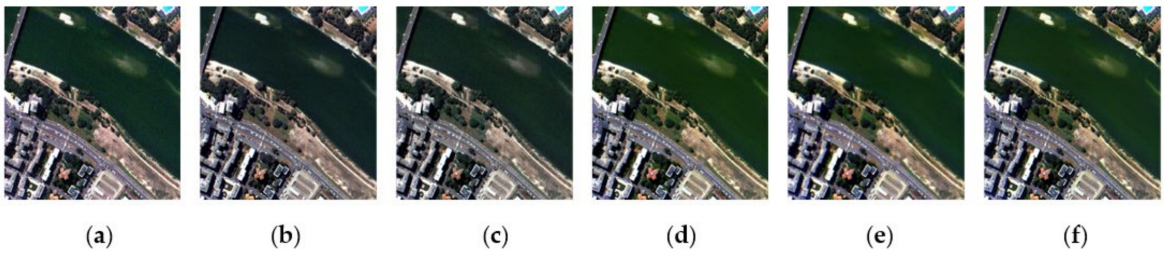

Figure 3. Fused images of Pavia City Center. (**a**) CNMF; (**b**) JNMF; (**c**) DPLM; (**d**) TCNNF; (**e**) DiCNN; (**f**) CpNN.

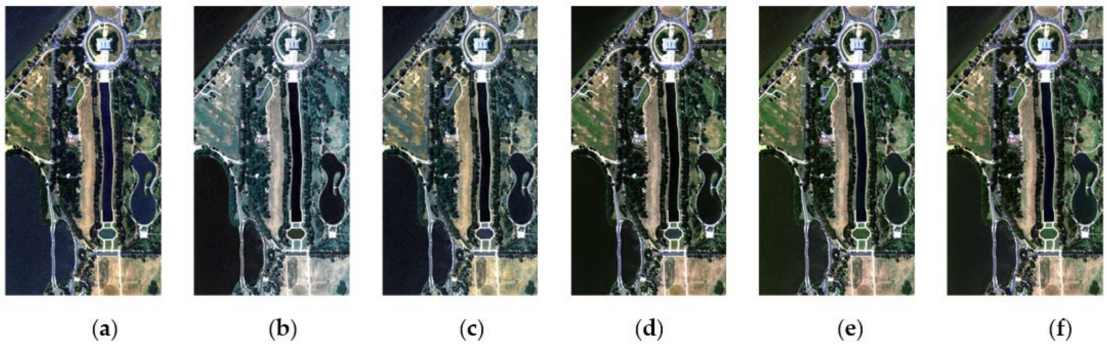

Figure 4. Fused images of Washington DC Mall. (**a**) CNMF; (**b**) JNMF; (**c**) DPLM; (**d**) TCNNF; (**e**) DiCNN; (**f**) CpCNN.

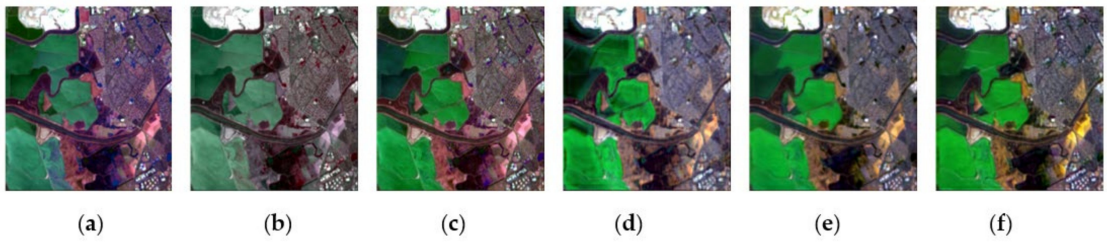

Figure 5. Fused images of San Francisco. (**a**) CNMF; (**b**) JNMF; (**c**) DPLM; (**d**) TCNNF; (**e**) DiCNN; (**f**) CpCNN.

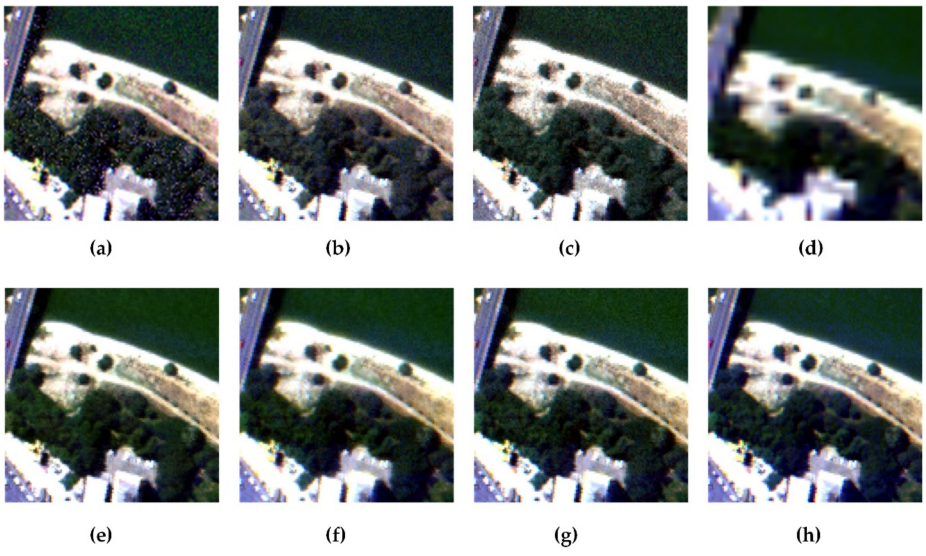

Figure 6. Local enlargements of fusion result of Pavia City center. (**a**) CNMF; (**b**) JNMF; (**c**) DPLM; (**d**) LR HS image; (**e**) TCNNF; (**f**) DiCNN; (**g**) CpCNN; (**h**) Reference high-resolution, hyperspectral (HR HS) image.

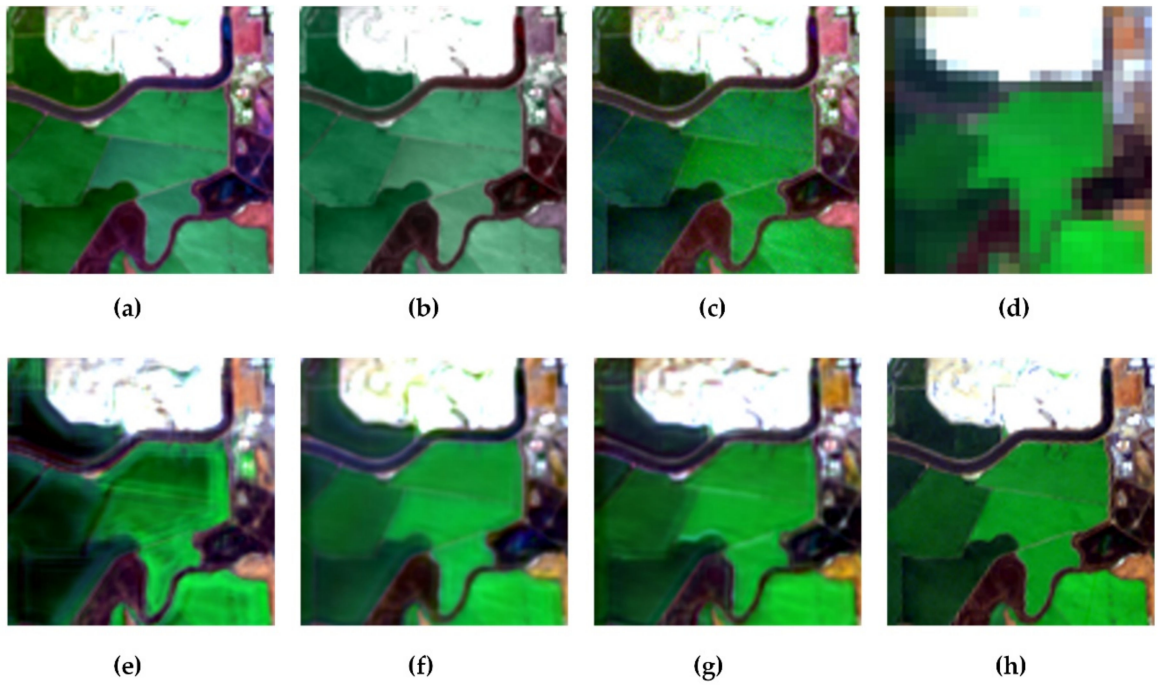

Figure 7. Local enlargements of fusion result of San Francisco. (**a**) CNMF; (**b**) JNMF; (**c**) DPLM; (**d**) LR HS image; (**e**) TCNNF; (**f**) DiCNN; (**g**) CpCNN; (**h**) Reference HR HS image.

3.3. Parameter Discussion

Finally, we would like to give a brief analysis of the parameters that may have certain effects on the fusion results in our presented method. Most parameters of our method were determined empirically according to the massive references, e.g., learning rate, batch size, number of epochs, and filter numbers. The 3×3 filter size is also suggested by numerous related works. Since the numbers of convolutional layers highly depend on the patch sizes in this network, we will focus on discussing two peculiar parameters, namely the patch sizes of HS and MS images. Figures 8 and 9 show the ERGAS and UIQI values of our method with respect to the patch sizes of HS and MS images, respectively. Due to the limitation of page length, only the results under a resolution ratio of 4 (i.e., $s = 4$) are plotted. However, similar results and conclusions can also be observed under different resolution ratios.

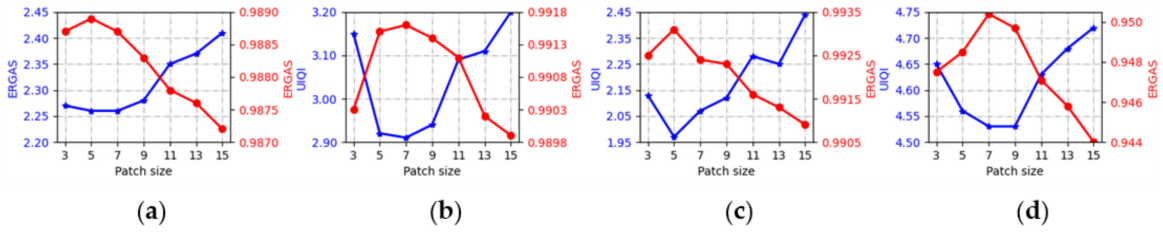

Figure 8. The relative dimensionless global error in synthesis (ERGAS) and universal image quality index (UIQI) values with respect to the patch size of HS image. (**a**) University of Pavia; (**b**) Pavia City center; (**c**) Washington DC Mall; (**d**) San Francisco.

As a matter of fact, pixels that are far from the center pixel scarcely affect the spectral and spatial details of the current pixel. Thus, a large local area is not necessary for predicting the features of this pixel. On the contrary, a larger patch size usually means that a more complicated architecture of network should be constructed to properly extract the structure information, which generally requires more training samples to obtain a stable network. Therefore, a proper size of patch deserves to be estimated. Meanwhile, it should

also be noted that the coupled network aims at exploiting the collaborative spectral and spatial features for HS and MS images, respectively. Nevertheless, the HS sub-network substantially focuses on approximating a spectrum for an unknown pixel according to its nearest neighbors, whereas the MS sub-network focuses on predicting the details by exploring the spatial correlations within a larger local area. As it is well-recognized that HS and MS image have different spatial resolutions, which means they substantially locate on different spatial scales from the viewpoint of multi-scale model, it is natural that using different sizes of patches will benefit the training of network.

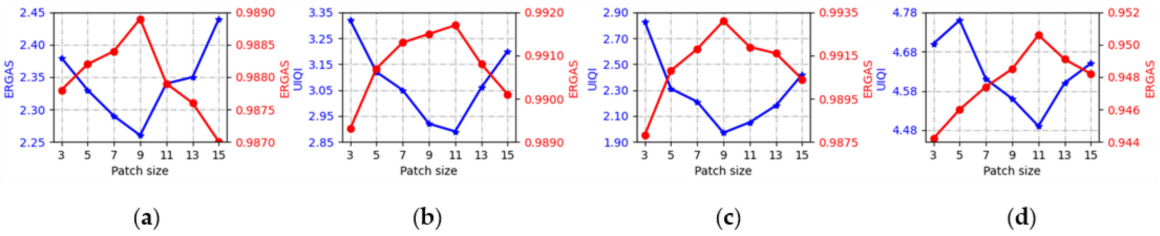

Figure 9. The ERGAS and UIQI values with respect to the patch size of MS image. (**a**) University of Pavia; (**b**) Pavia City center; (**c**) Washington DC Mall; (**d**) San Francisco.

From Figure 8, we can see that, generally, the patch size of 5×5 or 7×7 is a relatively proper choice for the HS sub-network. A larger patch size does not help improve the performance. Sizes larger than 11×11 are not recommended in our method. On the other hand, Figure 9 suggests that for the patch size of MS sub-network, the optimal results occur around 9×9 and 11×11. Since the MS image has higher spatial resolution than HS image, the patch size is commonly larger than that of the HS image in order to fully explore the spatial relations of neighboring pixels. Therefore, sizes smaller than 7×7 or larger than 13×13 are not recommended in our network. This is basically in accordance with the above theoretical analyses.

Consequently, in the experiments of Section 3.2., the patch sizes of HS and MS images are set to 5×5 and 9×9, respectively, without exception. As mentioned in Section 2.2., the two sub-networks, thus, contain 2 and 4 convolutional layers, respectively, without exception either.

4. Conclusions

Hyper-sharpening has attracted numerous research in the past two decades. Apart from the classical matrix factorization-based algorithms, convolutional neural networks show considerable potential in HS and MS image processing areas due to its adaptability and robustness of image feature extraction. To minimize the spectral distortion, in this paper, we propose to integrate the convolutional neural network with a detailed injection model for HS and MS image fusion. The proposed approach employs a couple of convolutional networks for feature extraction of HS and MS images individually and predicts the missing high-level spatial details. The network is concise and efficient and is able to achieve satisfactory performance in spectral fidelity aspect. Our future work will focus on the automatic selection of network parameters in order to further promote the flexibility and adaptability of the proposed method.

Author Contributions: Conceptualization, X.L., F.J.; data curation, X.L.; formal analysis, X.L., D.Y. and Y.Z.; funding acquisition, X.L. and F.J.; investigation, X.L.; methodology, X.L.; Project administration, F.J.; resources, D.Y.; software, D.Y.; supervision, F.J.; validation, X.L., D.Y. and Y.Z.; visualization, X.L. and Y.Z.; writing—original draft, X.L.; writing—review & editing, D.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Natural Science Foundation of Shanghai, grant number 19ZR1453800, and the Fundamental Research Funds for the Central Universities, grant number 2232020D-46.

Data Availability Statement: Data available in a publicly accessible repository that does not issue DOIs Publicly available datasets were analyzed in this study. Data can be found here: [http://www.ehu.eus/ccwintco/index.php/Hyperspectral_Remote_Sensing_Scenes; https://engineeringpurdue.edu/~biehl/MultiSpec/hyperspectral.html; https://earthexplorer.usgs.gov].

Acknowledgments: The authors would like to express their great appreciation to P. Gamba for providing the ROSIS data over Pavia, Italy, and Spectral Information Technology Application Center of Virginia for providing the Washington DC Mall data. The authors would also like to sincerely thank the United States Geological Survey (USGS) for their publicly available EO-1 Hyperion and ASTER data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Selva, M.; Aiazzi, B.; Butera, F.; Chiarantini, L.; Baronti, S. Hyper-sharpening: A first approach on SIM-GA data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 3008–3024. [CrossRef]
2. Chen, Z.; Pu, H.; Wang, B.; Jiang, G. Fusion of hyperspectral and multispectral images: A novel framework based on generalization of pan-sharpening methods. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1418–1422. [CrossRef]
3. Dong, W.; Liang, J.; Xiao, S. Saliency analysis and Gaussian mixture model-based detail extraction algorithm for hyperspectral pansharpening. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5462–5476. [CrossRef]
4. Li, X.; Yuan, Y.; Wang, Q. Hyperspectral and multispectral image fusion based on band simulation. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 479–483. [CrossRef]
5. Hardie, R.C.; Eismann, M.T.; Wilson, G.L. MAP estimation for hyperspectral image resolution enhancement using an auxiliary sensor. *IEEE Trans. Image Process.* **2004**, *13*, 1174–1184. [CrossRef]
6. Zhang, Y.; Backer, S.; Scheunders, P. Noised-resistant wavelet-based Bayesian fusion of multispectral and hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* **2009**, *47*, 3834–3843. [CrossRef]
7. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O.; Benediktsson, J.A. Model-based fusion of multi- and hyperspectral images using PCA and wavelets. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 2652–2663. [CrossRef]
8. Wei, Q.; Dobigeon, N.; Tourneret, J.Y. Fast fusion of multi-band images based on solving a Sylvester equation. *IEEE Trans. Image Process.* **2015**, *24*, 4109–4121. [CrossRef]
9. Wei, Q.; Dobigeon, N.; Tourneret, J.; Bioucas-Dias, J.; Godsill, S. R-FUSE: Robust fast fusion of multiband images based on solving a Sylvester equation. *IEEE Signal Process Lett.* **2016**, *23*, 1632–1636. [CrossRef]
10. Simões, M.; Bioucas-Dias, J.; Almeida, L.B.; Chanussot, J. A convex formulation for hyperspectral image superresolution via subspacebased regularization. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3373–3388. [CrossRef]
11. Wei, Q.; Bioucas-Dias, J.; Dobigeon, N.; Tourneret, J. Hyperspectral and multispectral image fusion based on a sparse representation. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3658–3668. [CrossRef]
12. Xing, C.; Wang, M.; Dong, C.; Duan, C.; Wang, Z. Joint sparse-collaborative representation to fuse hyperspectral and multispectral images. *Signal Process.* **2020**, *173*, 1–12. [CrossRef]
13. Yokoya, N.; Yairi, T.; Iwasaki, A. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 528–537. [CrossRef]
14. Song, H.; Huang, B.; Zhang, K.; Zhang, H. Spatio-spectral fusion of satellite images based on dictionary-pair learning. *Inf. Fusion.* **2014**, *18*, 148–160. [CrossRef]
15. Karoui, M.S.; Deville, Y.; Benhalouche, F.Z.; Boukerch, I. Hyper-sharpening by joint-criterion nonnegative matrix factorization. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 1660–1670. [CrossRef]
16. Lu, X.; Zhang, J.; Yu, X.; Tang, W.; Li, T.; Zhang, Y. Hyper-sharpening based on spectral modulation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 1534–1548. [CrossRef]
17. Borsoi, R.A.; Imbiriba, T.; Bermudez, J.C.M. Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability. *IEEE Trans. Image Process.* **2020**, *29*, 116–127. [CrossRef]
18. Li, S.; Dian, R.; Fang, L.; Bioucas-Dias, J.M. Fusing hyperspectral and multispectral images via coupled sparse tensor factorization. *IEEE Trans. Image Process.* **2018**, *27*, 4118–4130. [CrossRef]
19. Zhang, K.; Wang, M.; Yang, S.; Jiao, L. Spatial-spectral graph regularized low-rank tensor decomposition for multispectral and hyperspectral image fusion. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 1030–1040. [CrossRef]
20. Dian, R.; Li, S.; Fang, L. Learning a low tensor-train rank representation for hyperspectral image super-resolution. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2672–2683. [CrossRef]
21. Loncan, L.; De Almeida, L.B.; Bioucas-Dias, J.M.; Briottet, X.; Chanussot, J.; Dobigeon, N.; Liao, W.; Licciardi, G.A.; Simões, M.; Tourneret, J.-Y.; et al. Hyperspectral pansharpening: A Review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 27–46. [CrossRef]
22. Yokoya, N.; Grohnfeldt, C.; Chanussot, J. Hyperspectral and multispectral data fusion: A comparative review of the recent literature. *IEEE Geosci. Remote Sens.* **2017**, *5*, 29–56. [CrossRef]
23. Ahmad, M. A fast 3D CNN for hyperspectral image classification. *arXiv* **2020**, arXiv:2004.14152.
24. Yuan, Q.; Wei, Y.; Meng, X.; Shen, H.; Zhang, L. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2018**, *11*, 978–989. [CrossRef]

25. Masi, G.; Cozzolino, D.; Verdoliva, L.; Scarpa, G. Pansharpening by convolutional neural networks. *Remote Sens.* **2016**, *8*, 594. [CrossRef]
26. Kawulok, M.; Benecki, P.; Piechaczek, S.; Hrynczenko, K.; Kostrzewa, D.; Nalepa, J. Deep Learning for Multiple-Image Super-Resolution. *IEEE Geosci. Remote Sens. Lett.* **2020**, *17*, 1062–1066. [CrossRef]
27. Palsson, F.; Sveinsson, J.R.; Ulfarsson, M.O. Multispectral and Hyperspectral Image Fusion Using a 3-D-Convolutional Neural Network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 639–643. [CrossRef]
28. Dian, R.; Li, S.; Guo, A.; Fang, L. Deep hyperspectral image sharpening. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5345–5355. [CrossRef]
29. Yang, J.; Zhao, Q.; Chan, J.C. Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network. *Remote Sens.* **2018**, *10*, 800. [CrossRef]
30. Zhou, F.; Hang, R.; Liu, Q.; Yuan, X. Pyramid fully convolutional network for hyperspectral and multispectral image fusion. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 1549–1558. [CrossRef]
31. Dian, R.; Li, S.; Kang, X. Regularizing Hyperspectral and multispectral image fusion by CNN denoiser. *IEEE Trans Neural Netw. Learn Syst.* **2020**, 1–12. [CrossRef] [PubMed]
32. Wang, Z.; Chen, B.; Lu, R.; Zhang, H.; Liu, H.; Varshney, P.K. FusionNet: An unsupervised convolutional variational network for hyperspectral and multispectral image fusion. *IEEE Trans. Image Process.* **2020**, *29*, 7565–7577. [CrossRef]
33. Xu, S.; Amira, O.; Liu, J.; Zhang, C.X.; Zhang, J.; Li, G. HAM-MFN: Hyperspectral and multispectral image multiscale fusion network with RAP loss. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4618–4628. [CrossRef]
34. Shen, D.; Liu, J.; Xiao, Z.; Yang, J.; Xiao, L. A twice optimizing net with matrix decomposition for hyperspectral and multispectral image fusion. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2020**, *13*, 4095–4110. [CrossRef]
35. Lu, R.; Chen, B.; Cheng, Z.; Wang, P. RAFnet: Recurrent attention fusion network of hyperspectral and multispectral images. *Signal Process.* **2020**, *177*. [CrossRef]
36. Selva, M.; Santurri, L.; Baronti, S. Improving hypersharpening for WorldView-3 Data. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 987–991. [CrossRef]
37. He, L.; Rao, Y.; Li, J.; Chanussot, J.; Plaza, A.; Zhu, J.; Li, B. Pansharpening via detail injection based convolutional neural networks. *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.* **2019**, *12*, 1188–1204. [CrossRef]
38. Lu, X.; Zhang, J.; Li, T.; Zhang, Y. Pan-sharpening by multilevel interband structure modeling. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 892–896. [CrossRef]
39. Lu, X.; Zhang, J.; Li, T.; Zhang, Y. A novel synergetic classification approach for hyperspectral and panchromatic images based on self-learning. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 4917–4928. [CrossRef]