IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium | 978-1-6654-0369-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/IGARSS47720.2021.9554024

PROGRESSIVE BAND-SEPARATED CONVOLUTIONAL NEURAL NETWORK FOR MULTISPECTRAL PANSHARPENING

Shi-Shi Xiao[1], *Cheng Jin*[2], *Tian-Jing Zhang*[3], *Ran Ran*[4], *Liang-Jian Deng*[4,*]

[1]School of Information and Communication Engineering,
[2]School of Optoelectronic Science and Engineering,
[3]Yingcai Honors College, [4]School of Mathematical Sciences,
University of Electronic Science and Technology of China, Chengdu, 611731

ABSTRACT

Recently, convolutional neural networks (CNNs) have been introduced to pansharpening for enhancing fusion accuracy and overcoming the drawbacks of the conventional methods. However, most of methods based on CNN fail to distinguish the difference of multispectral bands, and only use a uniform set of convolutional kernels to extract features. In this paper, we design a progressive, band-separated convolutional network architecture for discriminatively learning the features and relation among spectral bands, aiming to address the problem mentioned before. More specifically, the proposed architecture mainly consists of three aspects. First, to accurately preserve the spectral peculiarities, we divide the multispectral input image in terms of its bands into several groups. Second, our original panchromatic and multispectral inputs are filtered by a high-pass operation to further yield more spatial details. Third, we use a spectral fusion module (SFM) for each group and associate them to progressively assemble the whole architecture. It is worth mentioning that the architecture could be integrated into any other competitive CNNs to improve the performance. Both visual and quantitative experiments have demonstrated that our proposed method outperforms recent state-of-the-art pansharpening techniques.

Index Terms— Band-Separated Convolutional Neural Network, Pansharpening, Multispectral Image, Progressive Network

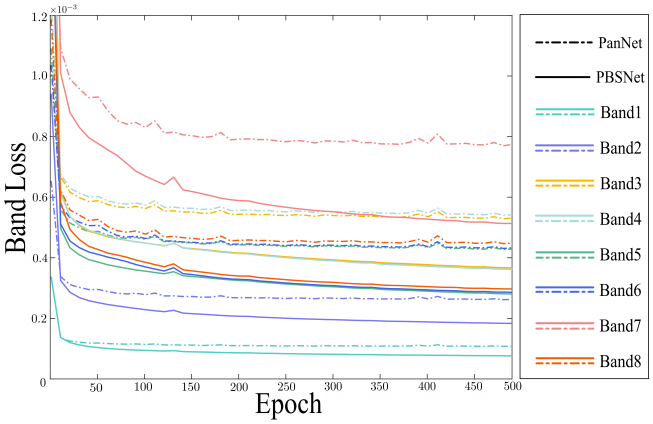

Fig. 1. Dotted lines: the training loss for each MS band obtained by PanNet; Solid lines: the training loss for each band by the proposed PBSNet. It is clear that PBSNet could significantly diminish the training loss for each MS band.

1. INTRODUCTION

Processing multispectral images is an elementary task with respect to remote sensing. Among techniques like spectral unmixing, image super-resolution, pansharpening is designated for producing multispectral images with high spatial resolution by merging high spatial resolution panchromatic (PAN) images with low spatial resolution multispectral (MS) images, which become available by virtue of multiple satellite sensors. As pansharpening has drawn much attention for decades, extensive algorithms have been raised constantly, for which we may classify into four mainstream methods: 1) component substitution (CS) [1]; 2) multi-resolution analysis (MRA) [2]; 3) variational optimization (VO) [3]; 4) deep learning (DL). The first three approaches may result in distortion in both spatial and spectral domain more or less, deteriorating the integral property of the fused image.

Thanks to massive real-world datasets and the development of hardware, the fourth approach based on DL is employed to break the bottleneck in pansharpening. Various innovative CNN structures are proposed to discern and extract hierarchical features at spatial resolution adaptively while spectral content simultaneously fused at high precision [4–6]. However, most contemporary methods adopt a uniform set of convolutional kernels to extract features regardless of the distinction of multispectral bands. To achieve more advanced performance of pansharpening, we propose a new architecture motivated by fully exploiting the different characteristics of each band. Bands of MS image are separated into several groups according to their proximity in training loss, then each group possesses an independent

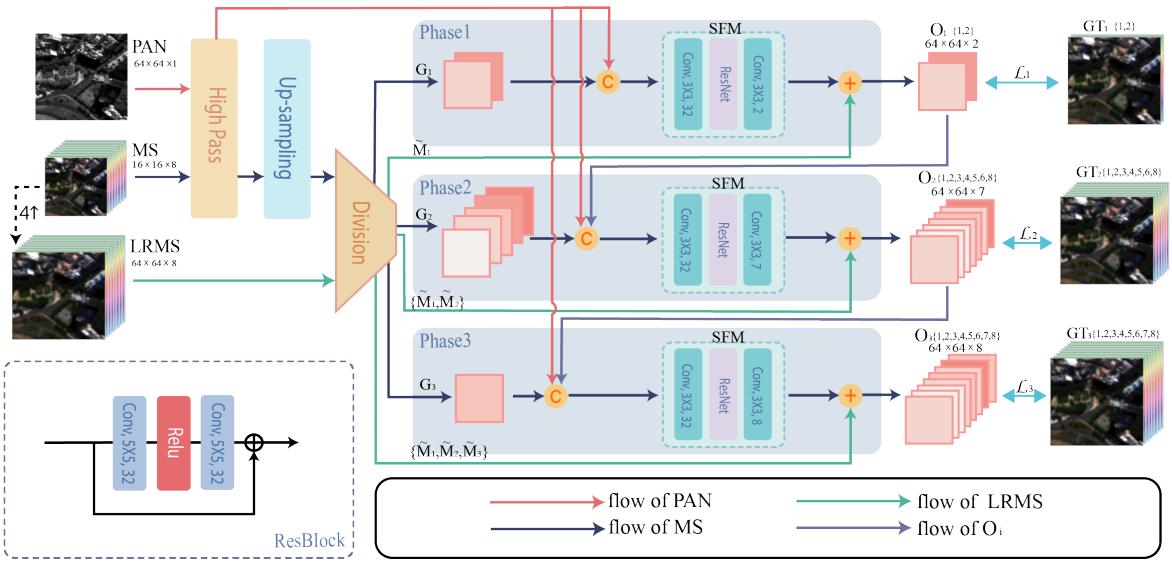

Fig. 2. The flowchart of our architecture (PBSNet) based on WorldView-3 dataset (8 bands). In which, the eight upsampled MS bands after high-pass filtering are divided into three groups, *i.e.*, $\boldsymbol{G}_1$ with the $1,2$-th bands, $\boldsymbol{G}_2$ with the $3,4,5,6,8$-th bands and $\boldsymbol{G}_3$ with the 7-th band. Also, $\tilde{\boldsymbol{M}}$ is divided with the same way to generate $\tilde{\boldsymbol{M}}_1$, $\tilde{\boldsymbol{M}}_2$ and $\tilde{\boldsymbol{M}}_3$. Especially, please notice the color of image flows (*i.e.*, arrows) shown in the bottom, it is quite crucial to understand the network architecture correctly.

but structurally identical module named SFM. Unlike existed methods such as SRPPNN [7] based on the stepwise upsampled scale, the SFM we proposed is integrated with the output of the last SFM to make the whole fusion architecture progressively, preserving inter-spectral content as much as possible. Furthermore, high-pass filters are employed to extract the high-frequency information from MS and PAN images, and then send it into the network to avoid quantitative deviation between the different satellites [4]. In summary, progressive band-separated network (PBSNet) has the following main contributions:

- Each SFM explicitly focuses on particular bands proximity in training loss, which efficiently exploits spatial features rather than the uniform convolutional kernels.
- The fusion performance is promised by integrating each SFM outputs progressively, aiding spectral loss of all bands and promoting spectral feature association.
- The proposed architecture can easily be integrated into any other competitive CNNs.
- Experimental results illustrate the superiority of PBSNet in comparison with other cutting-edge techniques.

2. THE PROPOSED METHOD

2.1. Motivation

Many algorithms based on CNN have recently developed and modified increasingly to a new level for better pansharpening performance. An evident fact for a multispectral image is that different spectral bands have quite distinct properties mainly because of the various wavelengths by sensors. Additionally, extensive experiments on CNNs methods provide us a key observation as Fig. 1 which shows that the outcomes of CNN-based approaches (*e.g.*, PanNet) hold significantly different training loss for each multispectral band due to the uniform set of convolutional kernels mentioned before. Thus, it motivates us to adopt a *divide-and-conquer* strategy and formulate a progressive band-separated CNN (*i.e.*, PBSNet) for multispectral pansharpening. Especially, Fig. 1 confirms the proposed PBSNet could significantly reduce the training loss of each multispectral band.

2.2. Overview of Network Architecture

Fig. 2 exhibits the overall network architecture of the proposed PBSNet. In the figure, $\boldsymbol{M} \in \mathbb{R}^{H\times W\times C}$ denotes the original MS image including C spectral bands with spatial size of $H \times W$; let $\tilde{\boldsymbol{M}} \in \mathbb{R}^{H\times W\times C}$ represent a upsampled MS image (LRMS) and $\boldsymbol{P} \in \mathbb{R}^{H\times W}$ be an observed PAN image; and $\boldsymbol{GT} \in \mathbb{R}^{H\times W\times C}$ refers to ground truth image.

Overall Structure To yield more spatial details, $\boldsymbol{M}$ and $\boldsymbol{P}$ are initially processed by high-pass filters, denoting as $\boldsymbol{M}_H$ and $\boldsymbol{P}_H$, respectively. By considering the proximity in training loss, the upsampled $\boldsymbol{M}_H$, *i.e.*, $\tilde{\boldsymbol{M}}_H$, is splited into several groups, denoted as $\boldsymbol{G}_n, n = 1, 2, \cdots, s$, where s is the total number of groups ($s = 3$ in this work for WV-3 data, see the following "Band Grouping" for more details).

Band Grouping In this work, we group the bands of a

multispectral image on the WorldView-3 (WV-3) dataset (8 MS bands) by analyzing the training loss shown in Fig. 1. It is clear that the $1, 2$-th, $3, 4, 5, 6, 8$-th and 7-th bands have abysmally different training losses due to the intrinsic properties of MS bands, which inspires us to categorize them into three groups. For simplification, we denote $\boldsymbol{G}_1$, $\boldsymbol{G}_2$ and $\boldsymbol{G}_3$ as the upsampled $\boldsymbol{M}_H$ bands with the corresponding group index. For other dataset, we also could do the grouping by the same strategy.

All inputs sent to SFM can be simply described as follows,

$$\begin{aligned} \boldsymbol{I}_1 &= \boldsymbol{\mathcal{C}}\left[\boldsymbol{G}_1, \boldsymbol{P}\right], \\ \boldsymbol{I}_n &= \boldsymbol{\mathcal{C}}\left[\boldsymbol{G}_n, \boldsymbol{P}, \boldsymbol{O}_{n-1}\right], n = 2, 3, \cdots, s, \end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{C}}$ stands for the concatenation operation, and $\boldsymbol{I}_n$ represents the network input consisting of $\boldsymbol{P}$, $\boldsymbol{G}_n$ and $\boldsymbol{O}_{n-1}$ (the output of the phase $n-1$). Especially, the output $\boldsymbol{O}_n$ on the phase n can be represented by the following formulas,

$$\boldsymbol{O}_n = \boldsymbol{\mathcal{F}}_n(\boldsymbol{I}_n) + \boldsymbol{\mathcal{C}}\left[\tilde{\boldsymbol{M}}_1, \cdots, \tilde{\boldsymbol{M}}_n\right], \tag{2}$$

where $\boldsymbol{\mathcal{F}}_n$ depicts a functional mapping between the input and output (also the SFM that will be introduced afterwards), $\tilde{\boldsymbol{M}}_n$ is the partial multispectral bands with the same group index as $\boldsymbol{G}_n$ (see more details from Fig. 2).

After defining the network architecture, we will define the final loss function for our PBSNet. Specifically, the loss function contains s losses to minimize the distance between the output and the GT images for each phase. Here, we employ ℓ_2 loss for better performance, see as follows,

$$\boldsymbol{\mathcal{L}}(\Theta) = \sum_{i=1}^{s} \boldsymbol{\mathcal{L}}_i(\Theta) = \sum_{i=1}^{s} \|\boldsymbol{O}_i - \boldsymbol{GT}_i\|_F^2, \tag{3}$$

where Θ indicates all parameters in this network, and $\boldsymbol{O}_i$ and $\boldsymbol{GT}_i$ represent the i-th network output and the ground truth (GT) image on the corresponding phase, respectively.

Spectral Fusion Module (SFM) As Fig. 2 shown, we formulate a simple SFM block mainly composed by a ResNet [8] with four ResBlocks. Within SFM, we apply an initial 3×3 convolution layer to obtain shallow features, and another 3×3 convolution layer to reconstruct the results. Hence, the SFM could deepen the network and extract more features for the fusion of pansharpened image.

3. RESULTS

In this section, the proposed PBSNet will be compared with some recent cutting-edge pansharpening approaches, including two traditional techniques (BDSD and GLP_CBD), two VO methods (CNMF, CVPR19), and three DL methods (DiCNN [9], PanNet [4] and DMDNet [5]).

Dataset and Metrics In this paper, training and testing of the DL methods are only based on WorldView-3 (WV-3) dataset. The corresponding training/ validation/ testing datasets are all the same as that in [10], in which the training dataset has 8806 samples generated by Wald protocol [11]. For the main network parameters of PBSNet, we set the learning rate as 1×10^{-3}, the batch-size as 32 and the total epochs as 500. Especially, we take the default parameters as the corresponding papers for the compared DL methods for fair comparisons. For the metrics on reduced-resolution test, we adopt the spectral angle mapper (SAM), the relative dimensionless global error in synthesis (ERGAS), the spatial correlation coefficient (SCC) and the universal image quality index for 8-band images (Q8). For the metric on full-resolution test, we utilize the quality with no inference (QNR), the spatial distortion index (D_λ) and the spectral distortion index (D_s). Please see more details from [10].

Table 1. Average quantitative results of different pansharpening approaches at 1258 reduced-resolution samples from a WV-3 dataset.

Method	SAM	ERGAS	SCC	Q8	Qave
BDSD	7.00±2.85	5.20±6.57	0.866±0.067	0.798±0.122	0.811±0.130
GLP_CBD	5.32±2.02	4.20±1.83	0.889±0.074	0.852±0.117	0.848±0.125
CNMF	6.06±1.27	4.80±1.48	0.866±0.104	0.836±0.142	0.855±0.123
SFIM	5.49±1.97	5.23±6.57	0.865±0.071	0.796±0.124	0.809±0.133
Reg_FS	5.26±1.94	4.16±1.77	0.891±0.069	0.854±0.115	0.850±0.123
PanNet	4.09±1.27	2.95±0.98	0.949±0.046	0.894±0.117	0.907±0.118
DiCNN	3.98±1.31	2.74±1.02	0.952±0.047	0.910±0.112	0.911±0.114
DMDNet	3.97±1.25	2.86±0.97	0.953±0.045	0.900±0.114	0.913±0.115
PBSNet	**3.58**±1.36	**2.43**±1.06	**0.962**±0.048	**0.921**±0.110	**0.920**±0.114

Table 2. Average quantitative results of different pansharpening approaches at 200 full-resolution samples from a WV-3 dataset.

Method	QNR	D_λ	D_s
EXP	0.920±0.011	0.000±0.000	0.080±0.100
BDSD	0.887±0.028	0.028±0.004	0.088±0.026
GLP_CBD	0.937±0.009	0.019±0.005	0.044±0.006
CNMF	0.930±0.016	0.028±0.010	0.044±0.009
CVPR19	0.941±0.010	0.008±0.002	0.051±0.010
PanNet	0.969±0.005	0.014±0.004	0.017±0.003
DiCNN	0.963±0.008	**0.013**±0.005	0.024±0.003
DMDNet	0.969±0.005	0.014±0.003	0.017±0.003
PBSNet	**0.974**±0.004	0.015±0.004	**0.012**±0.001

Reduced-resolution and Full-resolution Assessments Fig. 3 shows the visual comparisons of different techniques on reduced-resolution test. From this figure, the superiority of PBSNet can be noticed without evident artifacts and spectral distortion, and more edge information is sharpened, see the pansharpened images and residual maps in Fig. 3. Besides, the quantitative results reported in Tab. 1 also indicates that the PBSNet holds the best outcomes than other approaches on the average metrics of 1258 testing samples.

For the full-resolution test on 200 samples, the proposed method still outperforms other techniques on most of metrics

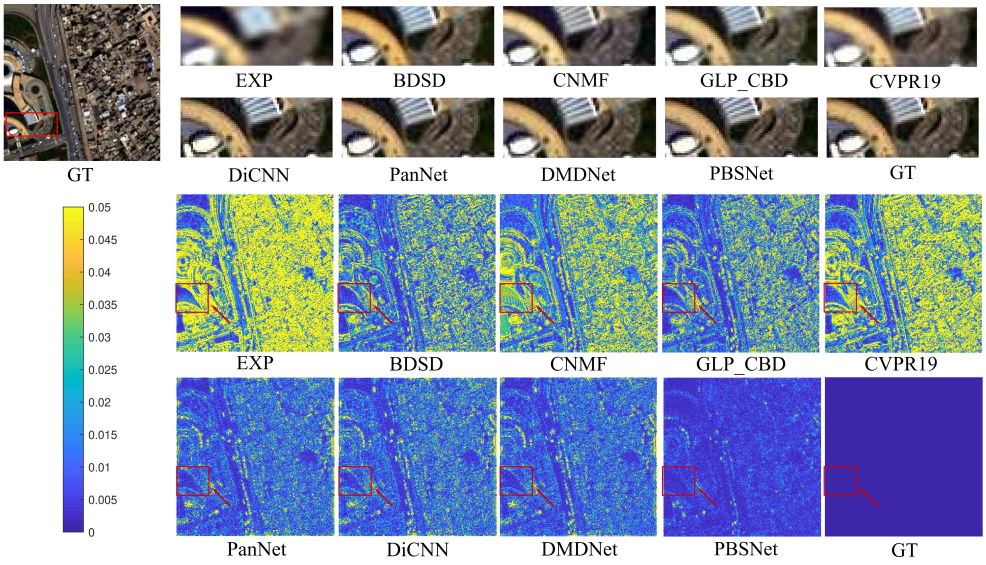

Fig. 3. The visual comparisons on a WorldView-3 data. The size of GT image is 256 × 256. First two rows: The fusion results by means of EXP, BDSD, CNMF, GLP_CBD, CVPR19, DiCNN, PanNet, DMDNet, and Proposed PBSNet. Third and fourth rows: The corresponding residual maps using GT image as reference. To aid the visual inspection, we display the residual maps obtained at the 4-th spectral band.

(especially the overall metrics QNR holding the first place), and achieves better spatial preservation and spectral fidelity, which also confirms the superior performance of PBSNet. Please see Tab. 2.

4. CONCLUSION

In this work, we introduced a progressive, band-separated convolution neural network for the multispectral pansharpening. This new network could discriminatively learn the features and relation among spectral bands, addressing the problem caused by using a uniform set of convolutional kernels. The progressive and band-separated architecture can remarkably to decrease spatial and spectral distortions. And the prominent performance of PBSNet was verified in broad visual and quantitative experiments on WV-3 dataset.

5. REFERENCES

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

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

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

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

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

[6] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.

[7] J. Cai and B. Huang, "Super-resolution-guided progressive pansharpening based on a deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, Aug. 2020, early access, doi: 10.1109/TGRS.2020.3015878.

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

[9] L. He, Y. Rao, J. Li, J. Chanussot, A. Plaza, J. Zhu, and B. Li, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Mar. 2019.

[10] L. Deng, G. Vivone, C. Jin, J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Trans. Geosci. Remote Sens.*, Oct. 2020, early access, doi: 10.1109/TGRS.2020.3031366.

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