# WEIGHTED SHALLOW-DEEP FEATURE FUSION NETWORK FOR PANSHARPENING

*Zi-Rong Jin, Tian-Jing Zhang, Cheng Jin, Liang-Jian Deng*

School of Optoelectronic Science and Engineering, University of Electronic Science and Technology of China

## ABSTRACT

In this paper, we propose a novel weighted shallow-deep feature fusion convolutional neural network (WSDFNet) for the task of multispectral image pansharpening. This network could effectively overcome the drawback of the common identity skip connection (ISC), and propagate shallow features scaled by a novel adaptive skip weighter (ASW) to deeper levels. By the technique, it could favor the feature fusion in different network depths adequately, as well as yield a promising outcome. Experimental results on reduced- and full-resolution WorldView-3 dataset demonstrate the superiority of the WSDFNet compared with recent state-of-the-art (SOTA) pansharpening approaches. Moreover, WSDFNet is also verified as a lightweight network.

***Index Terms*—** Pansharpening, Convolutional Neural Networks, Feature Fusion, Adaptive Skip Weighter

## 1. INTRODUCTION

Due to the rapid development of remote sensing image analysis, pansharpening has become a hot topic and attracted more attention in the scientific community. The goal of pansharpening is to get a high spatial resolution multispectral image by integrating high spatial resolution (HR) panchromatic (PAN) image and low spatial resolution (LR) multispectral (MS) image. At present, PAN and MS images are usually obtained from the ground scenes taken by remote sensing satellites, such as WorldView-3, QuikBird and GaoFen. Existing works to solve the pansharpening can be divided into four categories [1]: component substitution (CS) [2], [3] approaches, multi-resolution analysis (MRA) [4], [5] approaches, variational optimization (VO) [6] approaches, and deep learning (DL) approaches [7], [8]. Especially, CS and MRA based methods sometimes will lead to spatial and spectral distortions. This would bring the distortions at the interface of each operation. The methods based on the VO could get better performance through prior knowledge or some image degradation hypothesis. However, it is challenging to put forward reasonable assumptions and determine the appropriate priors.

Recently, with the gradual maturity of deep learning, the methods based on convolutional neural networks (CNNs) has also been applied to the field of pansharpening. Due to its powerful nonlinear and fitting abilities, the results obtained by these methods have been significantly improved over traditional methods in terms of evaluation indicators and visual perception. However, as most networks are designed based on the residual block with ISC (*e.g.*, PanNet [10], DMDNet [9]), there is a limitation for these networks to fuse the shallow features with the deep features, so that the shallow features are not fully utilized. In other words, the features extracted by the deep convolutional layers contain richer high-frequency information, if only simply stacking the residual blocks with ISC cannot effectively fuse the features extracted from shallow and deep convolutional layers. Thus, it motivates us to develop a weighted shallow-deep feature fusion network for pansharpening.

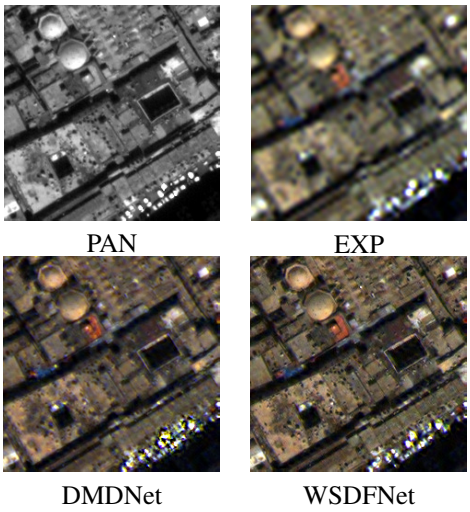


**Fig. 1**. The visual comparison on a full-resolution WorldView-3 dataset. First row: the original PAN image and the upsampled MS images (also called EXP). Second row: the pansharpened image by DMDNet [9] and WSDFNet.

In this paper, we adopt an adaptive skip weighter (ASW) to scale the shallow features for each residual block to fuse the features in different depths in the network to tackle this defect. The ASW could perform the feature fusion depending on the different inputs adaptively, which finally results in the so called WSDFNet.

The main contributions can be summarized as follows:

- A novel ASW module with a small amount of parame-

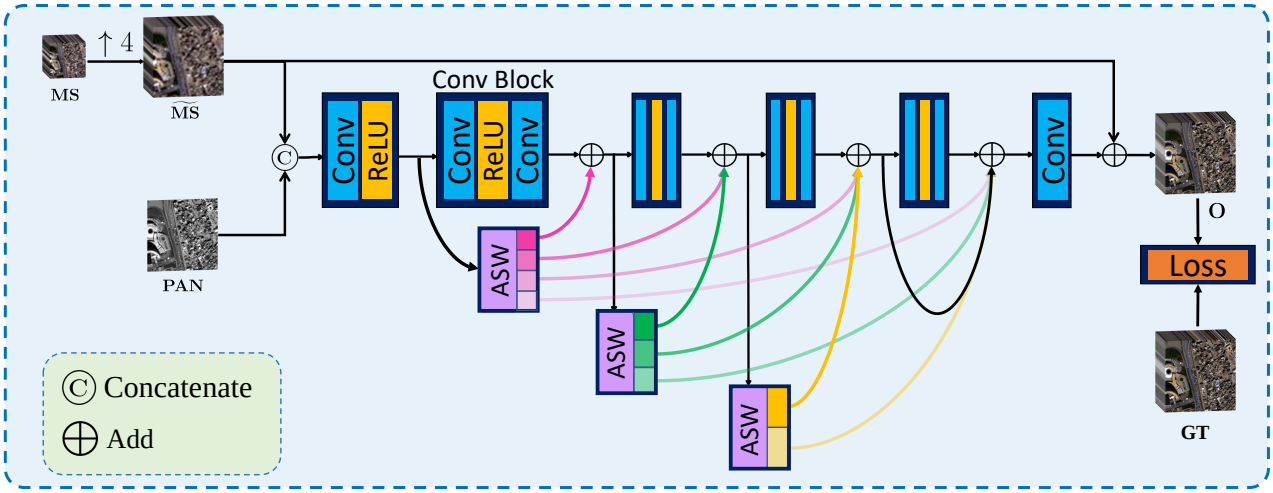


**Fig. 2**. The flowchart of WSDFNet. All involved convolution kernels in convolutional block (Conv Block) are with the size of $3 \times 3$ and 32 channels for simplicity. The related ASW and loss function can be found from Sec. 2.2 and Sec. 2.1.2.

ters is designed to adaptively scale the shallow features for each residual block so as to effectively fuse the features between shallow and deep layers.

- WSDFNet achieves the SOTA pansharpening performance, see Fig. 1. Especially, the given WSDFNet only involves about 80, 000 network parameters, holding a large gap compared with other DL-based methods, thus can be viewed as a lightweight network.

## 2. THE PROPOSED METHOD

In this section, WSDFNet will be stated in detail. Please note that in the following explanation, $\mathbf{P} \in \mathbb{R}^{H\times W\times 1}$ represents the PAN image where the $H$ and $W$ donates the size of $\mathbf{P}$ in spatial dimension, the $\mathbf{MS} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times b}$ represents the LRMS image where $b$ donates number of the multispectral bands, $\widetilde{\mathbf{MS}} \in \mathbb{R}^{H\times W\times b}$ is upsampled $\mathbf{MS}$ obtained by a polynomial kernel with 23 cofficients [11].

### 2.1. Network Architecture

#### 2.1.1. *Description of Network*

The overall architecture of WSDFNet is shown in Fig. 2, the $\widetilde{\mathbf{MS}}$ concatenates with $\mathbf{P}$ to get a concatenated image $\mathbf{I} \in \mathbb{R}^{H\times W\times(b+1)}$, which will be sent to WSDFNet. First, $\mathbf{I}$ will pass through a convolutional layer with the ReLU activation to extract the first shallow feature, which denoted as $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$, where $C$ donates the channels of the convolutional layer. After that, $\mathbf{X}$ will pass through several convolutianal blocks with ASW, the details of ASW can refer to Sec. 2.2. Finally, the output from the convolutional blocks will add with $\widetilde{\mathbf{MS}}$ to obtain the final output $\mathbf{O} \in \mathbb{R}^{H\times W\times b}$.

#### 2.1.2. *Loss Function*

To calculate the distance between the ground-truth (GT) image and $\mathbf{O}$, mean square error (MSE) is used as the loss function. It can be expressed as follows:

$$\mathcal{L}(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathcal{F}(\mathbf{MS}^{(i)}, \mathbf{P}^{(i)}; \Theta) - \mathbf{GT}^{(i)}\right\|_F^2, \quad (1)$$

where $\mathcal{F}(\cdot)$ stands for the given WSDFNet with the parameters $\Theta$, $N$ donates the amount of training examples, and $\|\cdot\|_F$ is the Frobenius norm.

### 2.2. Adaptive Skip Weighter

In this section, Fig. 3 shows the detailed structure of Adaptive Skip Weighter (ASW). From this figure, we consider an identity feature $\mathbf{X} \in \mathbb{R}^{H\times W\times C}$ as the input. A global average pooling (GAP) layer is first exploited to sample the feature of $\mathbf{X}$, then the sampled feature is sent to the following two sequential steps: 1) the first fully connected (FC) layer with a ReLU activation layer; 2) the second FC layer with a softmax activation layer. Finally, the obtained weight $\mathbf{W} \in \mathbb{R}^{1\times m}$ ($m$ represents the number of convolutional blocks) by the two FC layers is respectively multiplied by $\mathbf{X}$ to generate the weighted identity feature. Specifically, $\mathbf{W}$ could be expressed as follows,

$$\mathbf{W} = [w_1, w_2, \cdots, w_m], \quad (2)$$

where $w_i$ represents the skip weight scaling $\mathbf{X}$, and the obtained $w_i \cdot \mathbf{X}$ will be integrated to the $i^{th}$ convolutional block, see Fig. 2 for more details.

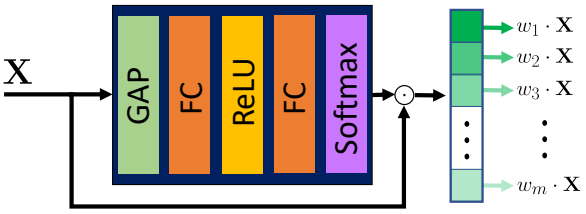


**Fig. 3**. The illustration of ASW.

## 3. RESULTS AND DISCUSSION

In this section, the proposed WSDFNet will compare with some recent SOTA pansharpening methods based on DL (*i.e.*, DiCNN1 [12], PanNet [10], DMDNet [9]) and some traditional approaches based on the CS (*i.e.*, BDSD [2], PRACS [3]) or the MRA (*i.e.*, CBD [4], SFIM [5]).

### 3.1. Dataset, Training details, and Parameters

All DL-based methods are fairly trained on the same dataset captured by WorldView-3 (WV3) satellite, as GT images are not available, thus Wald's protocol [13] is performed to ensure the baseline image generation. The original dataset can be download from the public website, and we obtain 12,580 PAN/MS/GT image pairs (70%/20%/10% as training/validation/testing dataset) with the size 64×64, 16×16×8, and 64×64×8 respectively by cropping the original dataset.

Besides, all DL-based methods are trained on NVIDIA GeForce GTX 2080Ti. For the parameters of our WSDFNet, we set the learning rate as $3 \times 10^{-4}$, epoch number as 500, and batch size as 32. For the compared approaches, we use the source code provided by the authors or re-implement the code with the default parameters in the corresponding papers.

### 3.2. Quality Assessment

To verify the effectiveness of our WSDFNet, we conduct the performance assessment on reduced resolution and full resolution. In the case of reduced resolution test, the relative dimensionless global error in synthesis (ERGAS), the spectral angle mapper (SAM), the spatial correlation coefficient (SCC), and 8-band images (Q8) are used to assess the quality of the results. In addition, to assess the performance of those methods on full resolutions, the QNR, the $D_\lambda$, and the $D_s$ indexes are applied. The experimental results are summarized in Table 1, Table 2 and Fig. 4.

Referring to the experimental results on reduced resolution and full resolution, it is evident that the proposed WSDFNet outperforms all the compared methods both in visual and quantitative results. In general, the DL-based methods perform significantly better than the traditional methods.

**Table 1**. Average quantitative comparisons on 1258 reduced resolution WorldView-3 examples.

| Method | *SAM* | *ERGAS* | *SCC* | *Q8* |
|---|---|---|---|---|
| **SFIM** | 5.452 ± 1.903 | 4.690 ± 6.574 | 0.866 ± 0.067 | 0.798 ± 0.122 |
| **PRACS** | 5.589 ± 2.981 | 5.167 ± 1.854 | 0.866 ± 0.081 | 0.813 ± 0.129 |
| **CBD** | 5.286 ± 1.958 | 4.163 ± 1.775 | 0.890 ± 0.070 | 0.854 ± 0.114 |
| **BDSD** | 7.000 ± 2.853 | 5.167 ± 2.248 | 0.871 ± 0.080 | 0.813 ± 0.123 |
| **DiCNN1** | 3.981 ± 1.318 | 2.737 ± 1.016 | 0.952 ± 0.047 | 0.910 ± 0.112 |
| **PanNet** | 4.092 ± 1.273 | 2.952 ± 0.978 | 0.949 ± 0.046 | 0.894 ± 0.117 |
| **DMDNet** | 3.971 ± 1.248 | 2.857 ± 0.966 | 0.953 ± 0.045 | 0.913 ± 0.115 |
| **WSDFNet** | **3.695 ± 1.226** | **2.544 ± 0.951** | **0.958 ± 0.046** | **0.915 ± 0.112** |
| **Ideal value** | **0** | **0** | **1** | **1** |

**Table 2**. Average quantitative comparisons on 25 full resolution WorldView-3 examples.

| Method | *QNR* | $D_\lambda$ | $D_s$ |
|---|---|---|---|
| **SFIM** | 0.934 ± 0.016 | 0.022 ± 0.008 | 0.045 ± 0.010 |
| **PRACS** | 0.907 ± 0.024 | 0.020 ± 0.006 | 0.075 ± 0.019 |
| **CBD** | 0.915 ± 0.024 | 0.028 ± 0.011 | 0.058 ± 0.015 |
| **BDSD** | 0.917 ± 0.029 | 0.020 ± 0.012 | 0.064 ± 0.022 |
| **DiCNN1** | 0.952 ± 0.012 | 0.017 ± 0.006 | 0.031 ± 0.008 |
| **PanNet** | 0.951 ± 0.010 | 0.031 ± 0.006 | **0.019 ± 0.005** |
| **DMDNet** | 0.944 ± 0.015 | 0.027 ± 0.007 | 0.031 ± 0.010 |
| **WSDFNet** | **0.964 ± 0.008** | **0.016 ± 0.005** | 0.020 ± 0.005 |
| **Ideal value** | **1** | **0** | **0** |

**Table 3**. Results for WSDFNet with or without ASW (WSDFNet w/o).

| Method | *SAM* | *ERGAS* | *SCC* | *Q8* |
|---|---|---|---|---|
| **WSDFNet w/o** | 2.946 | 1.855 | 0.965 | 0.970 |
| **WSDFNet** | **2.798** | **1.735** | **0.972** | **0.973** |

However, our WSDFNet has the best performance among these methods based on DL. Especially, the given WSDFNet also holds the smallest amount of parameters remarkably, *i.e.*, PanNet, DiCNN1, DMDNet and WSDFNet with the parameter amount of $2.5\times10^5$, $1.8\times10^5$, $3.2\times10^5$, $0.8\times10^5$, respectively.

### 3.3. The Validity of ASW

To verify the validity of ASW, we conduct an experiment by using the common ISC instead of the ASW while keeping all the other settings unchanged. From Table 3, the common ISC may weaken the original network performance for pansharpening, which confirms the effectiveness of ASW.

## 4. CONCLUSION

In this paper, a novel weighted shallow-deep feature fusion CNN (also called WSDFNet) was proposed for the multispectral image pansharpening. Different from the simple stack of residual blocks in general network structures, we used weighted identical skip connection to fuse the features extracted by the convolutional layers in different depths, and designed an adaptive skip weighter to dynamically adjust the importance of the features to the final fusion. It is worth mentioning that the proposed network is actually a lightweight

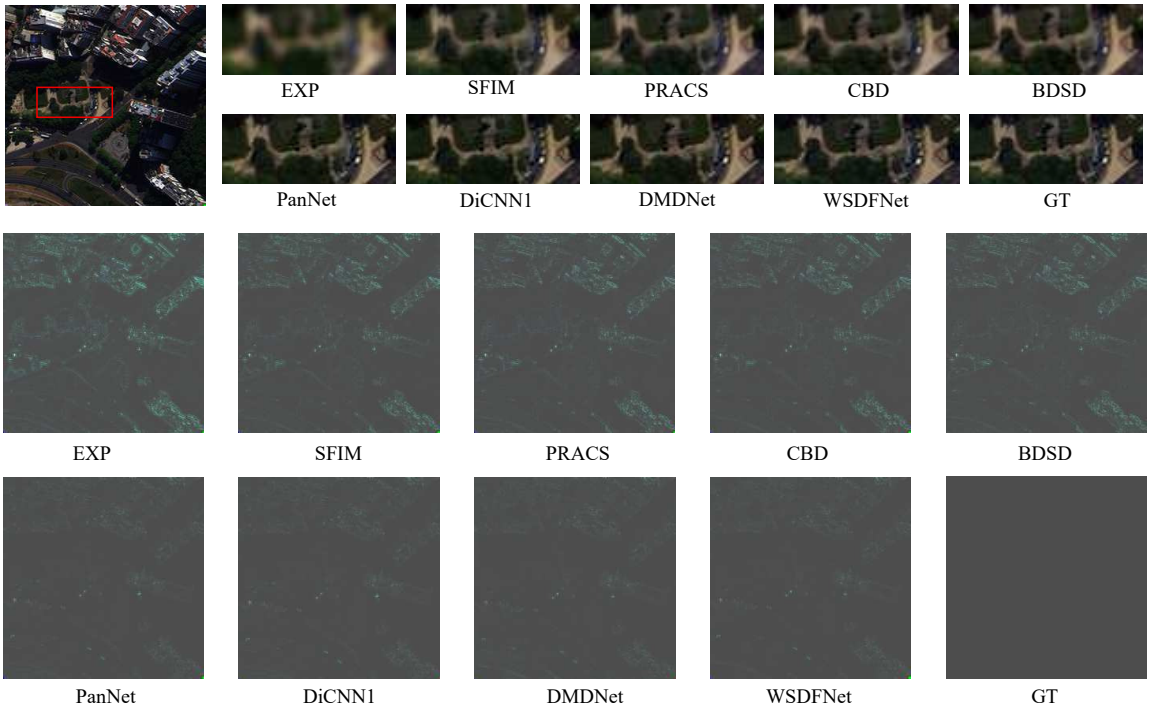


**Fig. 4**. First two rows: visual performance of compared approaches. Last two rows: the correponding residual maps.

one. Experimental results verify the mentioned contributions of the proposed WSDFNet.

## 5. REFERENCES


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

[2] A. Garzelli, F. Nencini, and L. Capobianco, "Optimal MMSE pan sharpening of very high resolution multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 1, pp. 228–236, 2008.

[3] 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, no. 1, pp. 295–309, 2010.

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

[5] J. Liu, "Smoothing filter-based intensity modulation: A spectral preserve image fusion technique for improving spatial details," *International Journal of Remote Sensing*, vol. 21, no. 18, pp. 3461–3472, 2000.

[6] 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 Transactions on Image Processing*, vol. 27, no. 9, pp. 4330–4344, 2018.

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

[8] L. Deng, G. Vivone, C. Jin, and J. Chanussot, "Detail injection-based deep convolutional neural networks for pansharpening," *IEEE Transactions on Geoscience and Remote Sensing*, 2020.

[9] 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.*, 2020, early access, doi: 10.1109/TNNLS.2020.2996498.

[10] 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.

[11] S. Baronti B. Aiazzi, L. Alparone 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, no. 10, pp. 2300–2312, 2002.

[12] 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, 2019.

[13] T. Ranchin and L. Wald, "Fusion of high spatial and spectral resolution images: The ARSIS concept and its implementation," *Photogramm. Eng. Remote Sensing*, vol. 66, no. 1, pp. 49–61, 2000.