COMPONENT SUBSTITUTION NETWORK FOR PAN-SHARPENING VIA SEMI-SUPERVISED LEARNING

Chi Liu, Yongjun Zhang*, Yangjun Ou

School of Remote Sensing and Information Engineering, Wuhan University,
Wuhan 430079, China - (liuchi, zhangyj, ouyangjun)@whu.edu.cn

Commission III, WG III/6

KEY WORDS: Pan-sharpening, component substitution, semi-supervised learning

ABSTRACT:

Pan-sharpening refers to the technology which fuses a low resolution multispectral image (MS) and a high resolution panchromatic (PAN) image into a high resolution multispectral image (HRMS). In this paper, we propose a Component Substitution Network (CSN) for pan-sharpening. By adding a feature exchange module (FEM) to the widely used encoder-decoder framework, we design a network following the general procedure of the traditional component substitution (CS) approaches. Encoder of the network decomposes the input image into spectral feature and structure feature. The FEM regroups the extracted features and combines the spectral feature of the MS image with the structure feature of the PAN image. The decoder is an inverse process of the encoder and reconstructs the image. The MS and the PAN image share the same encoder and decoder, which makes the network robust to spectral and spatial variations. To reduce the burden of data preparation and improve the performance on full-resolution data, the network is trained through semi-supervised learning with image patches at both reduced-resolution and full-resolution. Experiments performed on GeoEye-1 data verifies that the proposed network has achieved state-of-the-art performance, and the semi-supervised learning stategy further improves the performance on full-resolution data.

1. INTRODUCTION

Most optical remote sensing satellites provide both multispectral (MS) image and panchromatic (PAN) image. But due to hardware limitations, the MS image is of good spectral resolution while its spatial resolution is poor, and the PAN image is the other way around. Pan-sharpening aims at fusing the two to a synthetic image which is of good spectral resolution and spatial resolution. As an important means to improve data utilization of satellite images, pan-sharpening is often used as a preprocessing for other remote sensing tasks (Du et al., 2007, Mohammadzadeh et al., 2006).

The key issue of pan-sharpening is how to increase the spatial resolution of the MS image without introducing changes in spectral characteristics. In the last decades, different methods have been proposed to address the problem. Traditional approaches are component-substitution (CS) methods and multiresolution analysis (MRA) methods. The CS methods transform the MS image into a new domain in which one of the component is substituted by the PAN image, and reconstruct the image with an inverse transformation. Representative algorithms are principal component analysis (PCA) (Kwarteng, Chavez, 1989), intensity hue saturation (IHS) transform (Tu et al., 2004), and Gram-Schmidt (GS) sharpening (Laben, Brower, 2000). However, there are significant differences in spectral characteristics between the MS image and the PAN image, making the CS methods suffer from spectral distortion. The MRA methods extract multiscale details from the PAN image and inject them into the MS image. Representative algorithms are Laplacian pyramid (Aiazzi et al., 2002), wavelet transform (Nunez et al., 1999), curvelets transform (Nencini et al., 2007). But for the MRA methods, the quality of the output images is sensitive to the details injected. Insufficient details injection leads to blurring effects and excessive details injection results in artifacts and spectral distortions. During last decade, a series of model optimization (MO) methods (Ghahremani, Ghassemian, 2016, Fasbender et al., 2008, Palsson et al., 2014) have emerged, these methods model the relationship among the MS image, the PAN image and the desired high resolution multispectral (HRMS) image based on some reasonable assumptions and solve the model with some regularizations or priori constraints. However, proper models are difficult to build, and the model solving is time-consuming.

In recent years, deep learning has been developing rapidly and being widely used in various fields. Many researches (Ledig et al., 2017, Lin et al., 2017, Nah et al., 2017) have verified that deep learning networks are extremely suitable for computer vision tasks and have achieved state-of-the-art performance. As a typical low-level computer vision task, pan-sharpening also benefits from deep learning and many pan-sharpening networks have been proposed. The networks proposed earlier learn from single image super resolution (SISR) task, network structures such as sparse denoising auto encoders networks (Huang et al., 2015) and deep residual convolutional networks (Wei, Yuan, 2017) are used. By using the downsampled PAN image and the PAN image as input and output respectively, the networks learn a mapping from low resolution image to high resolution image. However, pan-sharpening is different from SISR because the details are extracted from the PAN image rather than inferred from the low resolution image. In these methods, the networks are trained on PAN image but directly applied for the MS image, the quality of the output image cannot be fully guaranteed. Then, some networks concatenate the upsampled MS image with the PAN image to form a synthetic image, which is used as input of the networks. Output of the networks is the pan-sharpened image. Masi et al. (Masi et al., 2016) proposed a shallow network with only three convolutional layers for pan-

*Corresponding author

sharpening. Scarpa et al. (Scarpa et al., 2018) further improved the network by using a deeper network with residual-learning and adding a target-adaptive tuning phase. Wei et al. (Wei et al., 2017) used an 11-layer deep residual network and adopted convolution kernels of larger size for better performance. Yuan et al (Yuan et al., 2018) proposed a two-stream network and used convolutional kernels of different sizes to extract multiscale features. Though these networks can output the pan-sharpened image end to end, the theoretical supports of these networks are scarce, it is difficult to explain how the networks handling the pan-sharpening task. More recently, several details injection networks have been proposed. He et al. (He et al., 2019) proposed two details injection networks, both the networks used three convolutional layers to get the residual image between the MS image and the pan-sharpened image. Zhang et al. (Zhang et al., 2019) used a bidirectional pyramid network to extract multiscale details from the PAN image and inject them into the MS image. Li et al. (Li et al., 2019) adopted a super-resolution network to upsample the MS image and then used guided filter to get the pan-sharpened image.

Though deep networks have shown great potential in pan-sharpening, there is a common drawback to existing networks. The training samples for pan-sharpening are generated by downsampling the original image. In another word, the networks are trained on images at reduced-resolution. It is often the case that a trained model performs well on reduced-resolution data but badly on full-resolution data. This drawback greatly limits the networks for practical applications, where the inputs are full-resolution images.

In this paper, we design a component substitution network for pan-sharpening. The network takes advantages of convolutional network and overcomes the drawbacks of traditional CS methods. The main contributions of this paper can be summarized as follows:

1. Following the general procedure of the CS approaches, the network extracts spectral feature and structure feature from the MS image and the PAN respectively, and uses a feature exchange module (FEM) to regroup the features.
2. To improve the performance on full-resolution data, we adopt a semi-supervised learning strategy. Besides reduced-resolution images, full-resolution images are also used in the training to improve the performance on full-resolution data.

The remainder of this paper is organized as follows. Section II summarizes the general process of CS methods and briefly introduces the semi-supervised learning. A detailed description of the proposed network is presented in Section III. The experimental results and assessments are presented and discussed in Section IV. Finally, a discussion and the conclusion are given in Section V.

2. RELATED WORK

2.1 Component Substitution Methods

Due to their impressive spatial quality and to their low computational cost, CS techniques have been widely investigated by the research community. The CS methods are based on the projection of the MS image into another space, assuming that the transformation separates the spatial structure from the spectral information in different components. The fusion process of the CS methods can be divided into 3 steps: 1). calculating the intensity component 2). component substitution 3). inverse transformation. A faster implementation of CS methods is formulated by:

$$\widehat{MS}_b = \widetilde{MS}_b + g_b(PAN - I), b = 1, ..., B \quad (1)$$

in which the subscript b is the band index, g_b are the inject gains, $\widetilde{MS}$ is the upsampled MS image and $\widehat{MS}$ is the pan-sharpened image. I is the intensity component and is calculated by:

$$I = \sum_{b=1}^{B} w_b * \widetilde{MS}_b \quad (2)$$

Different CS methods are mainly different in the calculation of w_b and g_b, but they're essentially linear calculations. The transformations can not completely separate the spatial structure from the spectral information, and the injection process inevitablely introduces spectral distortion.

2.2 Semi-Supervised Learning

Creating large datasets (Deng et al., 2009, Lin et al., 2014) typically requires a great deal of human effort. In many cases, it is difficult to obtain sufficient labeled data to train an effective model. Semi-supervised learning is an attractive approach towards addressing this problem. Based on the continuity and consistency in distribution between the labeled data and the unlabeled data, it attempts to automatically exploit unlabeled data in addition to labeled data to improve learning performance and enhance the representational ability of the model.

In the existing semi-supervised learning approaches, self-training (Rosenberg et al., 2005) is the simplest one. It starts by training on the labeled data only. In each step a part of the unlabeled data is labeled according to the current decision function; then the supervised method is retrained using its own predictions as additional labeled data. Another group of semi-supervised learning approaches are the graph-based methods (Blum et al., 2004, Zhu et al., 2003). These methods aim at constructing a graph connecting similar observations; label information propagates through the graph from labelled to unlabeled nodes by finding the minimum energy (MAP) configuration. There are also neural network-based approaches which combine unsupervised and supervised learning by training feed-forward classifiers with an additional penalty from an auto-encoder or other unsupervised embedding of the data (Ranzato, Szummer, 2008, Weston et al., 2012).

Recently due to the great advances of deep learning, semi-supervised processes have been applied successfully in more areas. Nasim Souly et al. (Souly et al., 2017) trained a generative adversarial networks (GANs) for semantic segmentation with semi-supervised learning. Yevhen Kuznietsov et al. (Kuznietsov et al., 2017) used semi-supervised learning to predict depth map from monocular images. Yong Cheng (Cheng, 2019) proposed a semi-supervised approach for training neural machine translation models on the concatenation of parallel corpora data and monolingual corpora data.

3. METHODOLOGY

3.1 framework

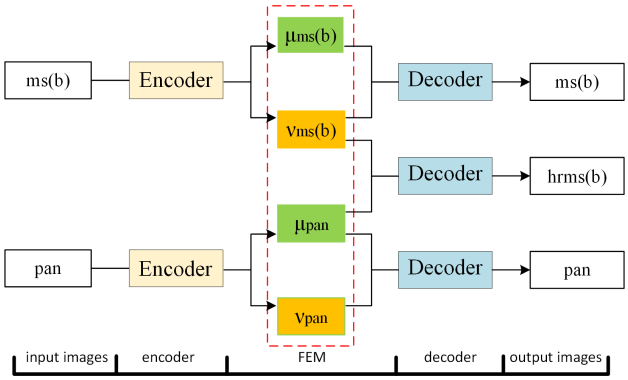


Figure 1. Framework of the proposed component substitution network. The FEM is inside the red dotted region.

Fig. 1 shows the overall structure of the proposed network. Just like the traditional CS methods, our network is composed of three parts, i.e., the encoder, the FEM, and the decoder. The encoder is corresponding to the band-transformation part in CS methods, it transforms the input image to a feature space in which the spectral information and the spatial information are separated. The FEM is corresponding to the CS process, it regroups the encoded features and combines the spectral feature of the MS image with the structure feature of the PAN image. The decoder is corresponding to the inverse transformation in CS methods, it reconstructs the image from the feature space.

As mentioned earlier, traditional CS methods tend to suffer from spectral distortion. The reason is that the spectral characteristics of the MS image and the PAN image are quite different, so the substituted component, which is the weighted sum of MS bands, is also different from the PAN image in spectral characteristics. Though strategies like histogram matching and local coefficient calculation are adopted to reduce the difference, spectral distortions are inevitably introduced. Deep learning gives solution to this problem. The high nonlinearity of the convolutional neural network enables it to explore high level information and find a feature space in which the spectral feature and the spatial feature are better separated.

Another drawback of traditional CS methods is that it ignores the spatial difference between different bands of the MS image. They use a single band component to represent the spatial characteristics of all the band of the MS image. Though the spatial resolution of each band of the MS image is the same, their spatial characteristics are not exactly the same. An example is given in Fig .2. In the images of the blue band and the red band, the boundary of the road and the vegetation in region 1 is invisible. In the image of the green band, the blue building in region 2 is mixed with the grass land. In the image of the near-infrared band, the grass land and the bare land in region 3 is mixed, and the buildings in region 4 is invisible. These differences demonstrate that it is unreasonable to extract the same spatial information for different bands of the MS image.

To address this problem, our network adopts single band images as input and processes the MS image band by band. The encoder extracts the spectral characteristic and spatial characteristic of a single band image rather than the MS image. As spectral characteristic and spatial characteristic are two basic aspects of remote sensing images, the same encoder can also be used by the PAN image. Thus, in the proposed network, we also use the same encoder for the PAN image and different bands of the MS image. The encoding process can be formulated as:

$$\mu_{ms}(b), \nu_{ms}(b) = encoder(ms(b)) \tag{3}$$

$$\mu_{pan}, \nu_{pan} = encoder(pan) \tag{4}$$

where b is the band index, $\mu_{ms}(b)$ and $\nu_{ms}(b)$ are the spatial feature and the spectral feature of the b_{th} band in the input MS image, respectively; μ_{pan} and ν_{pan} are the spatial feature and the spectral feature of the input PAN image, respectively.

In the FEM, the spectral feature of the MS image is regrouped with the structure feature of the PAN image. Then the regrouped features are sent to the decoder to get the pan-sharpened image, which can be formulated as:

$$\widehat{MS}_g(b) = decoder(\nu_{ms}(b), \mu_{pan}) \tag{5}$$

where $\widehat{MS}_g(b)$ is the b_{th} band of the reconstructed HRMS image.

The transformation in CS method is reversible, and the transformed components can fully recover the original image by an inverse-transformation. This guarantees there is no information loss in the transformation so that the components contain all the information of the original image. Similarly, in the proposed method, the decoder should also be able to recover the input image from the encoder. That is to say, the decoder should be able to recover the MS image with μ_{ms} and ν_{ms}, and be able to recover the PAN image with μ_{pan} and ν_{pan}. This process can be formulated as:

$$ms_g(b) = decoder(\mu_{ms}(b), \nu_{ms}(b)) \tag{6}$$

$$pan_g = decoder(\mu_{pan}, \nu_{pan}) \tag{7}$$

where $ms_g(b)$ is the b_{th} band of the reconstructed MS image, pan_g is the reconstructed PAN image.

Parameters of the proposed network is shared in two levels.

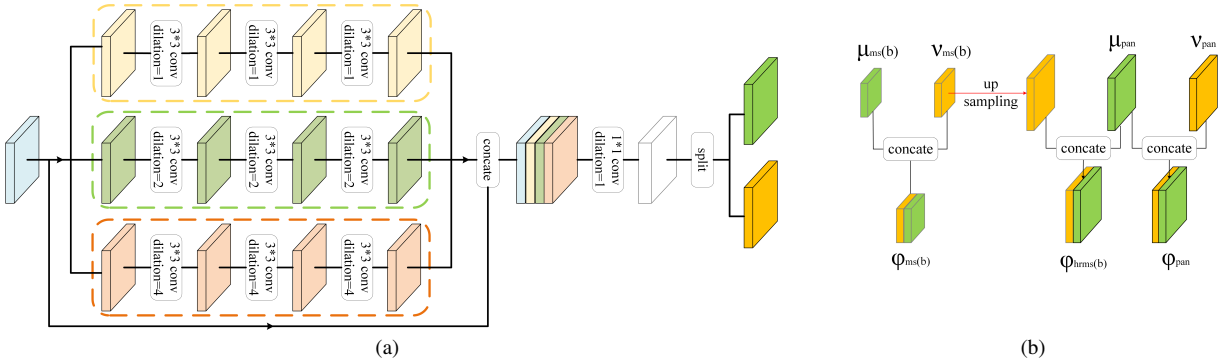
- Network level: The network processes the MS image band by band. Each band of the MS image together with the PAN image compose a pair of input image for the network. The network is shared by the image pairs. In the training, each band of the MS image is used as a training sample, and much fewer training image pairs are needed. This also permits the network to handle MS images with any number of bands.
- Module level: Both the encoder and the decoder of the network are shared by different inputs. The same encoder is used by the PAN image and each band of the MS image, and this enforces the encoder to be robust to spectral and spatial variations. The same decoder is used to reconstruct the MS image, the PAN image and the pan-sharpened image, and this ensures that there is no loss of information throughout the process.

3.2 Structure Details

Pan-sharpening is a task closely related to spatial resolution, so multiscale analysis has to be taken into consideration. In

(a) (b) (c) (d) (e)

Figure 2. An example of spatial difference between different bands of the MS image. From (a) to (d) are the blue band, the green band, the red band and the near-infared band, respectively. (e) is the true color MS image. Some typical regions are marked with a red box.



(a) (b)

Figure 3. Detail structure of the encoder and the FEM. (a) is the encoder. (b) is the FEM.

the widely used encoder-decoder framework, multiscale analysis is usually accomplished through pooling layers and deconvolutional layers. The networks downsample the feature maps in the encoder and unsample the feature maps in the decoder. But for the pan-sharpening task, the training samples have already been downsampled. Taking the GE1 image as an example, the resolution of training samples is 1/4 that of the original image, and after two pooling layers, the scale factor reduction compared to the original image will be 1/16. Such downsampling-upsampling operations lead to a serious loss of spatial information. Instead of downsampling the feature maps, we use atrous convolutional layers (Chen et al., 2017) to extract multiscale features and maintain the resolution. With different dilation rates, the kernels have different receptive fields to extract features at different scales. As shown in Fig. 3(a), the first branch uses convolutional kernels with dilation rate = 1, it actually uses ordinary convolutional layer and extracts local information. The second branch and the third branch use convolutional kernels with dilation rate = 2 and 4, respectively, they use convolutional layers with large receptive filed to extracts regional information. As pan-sharpening is a low-level task which concentrates on local region, we do not use convolutional kernels with larger dilation rate.

As the output of the encoder is of the same size with the input image, the feature maps of the MS image are different from that of the PAN image. So in the FEM, the spectral feature map of the MS image is upsampled to match the size of the PAN image. We simply use a bilinear interpolation to upsample the feature map because the spectral feature maps should be spatially smooth, and this also avoids extra computation brought in by a deconvolutional layer. The structure of the FEM is shown in Fig. 3(b), $\psi_{ms}(b)$, $\psi_{hrms}(b)$, and ψ_{pan} are the regrouped features, and they are sent to the decoder to reconstruct the single band MS image, the single band HRMS image and the PAN image, respectively. Since the encoded features have already considered multiscale analysis, we simply use three stacked ResBlocks to fuse the features. At the end of the decoder, a convolutional layer and a Relu layer are used to reconstruct the single band image from the feature map.

3.3 Network training with semi-supervised learning

For the existing pan-sharpening networks, the training samples are obtained by downsampling the full-resolution images. Downsampled MS image and downsampled PAN image are used as inputs to the network, and the original MS image is used as ground truth for the pan-sharpened image. By this mean, the burden of data preparation is greatly reduced. However, uncertainty has also been introduced. The spatial resolution of the training samples is much coarser than the full-resolution images, which results in decline in performance for full-resolution data. That's why some deep learning based methods do well on data at reduced-resolution but poor on data at full-resolution.

To improve the performance on full-resolution data, we use semi-supervised learning to train the network. Besides using reduced-resolution images as supervised training data, we also use full-resolution image as unsupervised training data. Illustration of the training process is shown in Fig. 4. $loss_r$ is the supervised training loss, and can be formulated as:

$$loss_r = loss_{ms} + loss_{pan} + loss_{\widehat{MS}} \tag{8}$$

where $loss_{ms}$ and $loss_{pan}$ measure the difference between the input and output MS and PAN image, respectively. $loss_{\widehat{MS}}$ measures the difference between the output HRMS and the

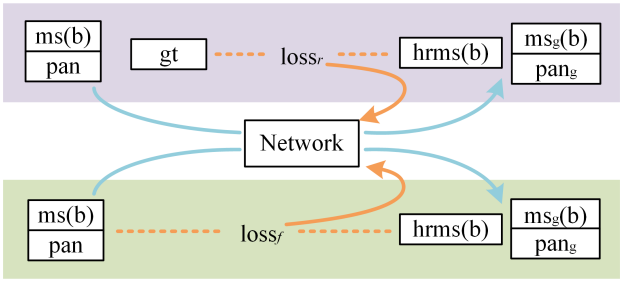


Figure 4. Training process of the proposed semi-supervised learning. The upper part is the supervised training with reduced-resolution data, the lower part is the unsupervised training with full-resolution data. The blue lines are the forward propagation while the orange lines are the backward propagation.

ground truth image. They are calculated by:

$$loss_{ms} = \|ms(b) - ms_g(b)\| \tag{9}$$

$$loss_{pan} = \|pan - pan_g\| \tag{10}$$

$$loss_{\widehat{MS}} = \|gt(b) - \widehat{MS}_g(b)\| \tag{11}$$

where $\|\ \|$ is the l_2 loss; $ms(b)$, pan and $gt(b)$ are the input MS image, the input PAN image and the ground truth pan-sharpened image, respectively; $ms_g(b)$, pan_g and $\widehat{MS}_g(b)$ are the reconstructed MS image, the reconstructed PAN image and the output pan-sharpened image, respectively. $loss_f$ is the unsupervised training loss, and can be formulated as:

$$loss_f = loss_{ms} + loss_{pan} + loss_{h2m} + loss_{h2p} \tag{12}$$

The first two term is the same as eq.8. $loss_{h2m}$ measures the spectral difference between the input MS image and the output pan-sharpened image, and it is based on the assumption that the degraded pan-sharpened image should be as similar as possible to the MS image. $loss_{h2p}$ measures the spatial difference between the input PAN image and the output pan-sharpened image. The two losses are calculated by:

$$loss_{h2m} = \|\widehat{MS}_g(b)\downarrow - ms(b)\| \tag{13}$$

$$loss_{h2p} = CC(\widehat{MS}_g(b) - \widetilde{MS}_g(b), pan - \widetilde{pan}\downarrow) \tag{14}$$

where $\downarrow$ is the downsampling operation, it downsamples the image according to the MTF of the sensor; CC is the correlation coefficient.

The full-resolution data used in training not only improves the performance in training, but also further alleviates the burden of data preparation. The acquirement of full-resolution data is easier compared to the reduced-resolution data, and much fewer images is needed to produce the same number of training samples.

4. EXPERIMENT

To evaluate the effectiveness of the proposed network and the semi-supervised learning strategy, we have conducted a series of experiments on both full-resolution data and reduced-resolution data. The results are compared by some objective indexes as well as visual appearance.

4.1 Experimental setup

We experiment with images acquired by GeoEye-1(GE1) with scenes concerning various areas. Following Wald's protocol, spatially degraded images with 8m resolution were used as inputs, and the original MS images with 2m resolution were used as the reference images. Totally, 1000 down-sampled image pairs were prepared, in which 800 pairs were used for training and 200 pairs were used for testing. In addition, 5500 full-resolution image pairs were prepared, in which 4000 were used for semi-supervised training and 1500 were used for testing. The patch size was 100*100 pixels for the MS image and 400*400 for the PAN image and the reference image. The network was optimized using Adam optimization algorithm based on back propagation, the learning rate was set to $1 * e^{-4}$.

To verify the effectiveness of the semi-supervised learning strategy, we trained two different models. The first one was only trained on the reduced-resolution training set and is called CSN, the other one was trained on both reduced-resolution data and full-resolution data and is called CSN-SSL. Besides our proposed method, six state-of-the-art pan-sharpening methods are used for comparison, i.e., BDSD (Garzelli et al., 2008), MMP (Kang et al., 2014), MTF-GLP (Aiazzi et al., 2002), GLP-SEGM (Restaino et al., 2017), PNN (Scarpa et al., 2018), MSDCNN (Yuan et al., 2018). BDSD and MMP are two CS methods, MTF-GLP and GLP-SEGM are two MRA methods, PNN and MSDCNN are two deep learning-based approaches. To make a comprehensive assessment, test results of different methods are evaluated by a series of indexes, both at reduced-resolution and full-resolution. Five indexes are chose for reduced-resolution evaluation, i.e., correlation coefficient (CC), relative dimensionless global error in synthesis (ERGAS) (Wald, 2002), root-mean-square error (RMSE), spectral angle mapper (SAM) (Yuhas et al., 1992) and structural similarity index (SSIM) (Wang et al., 2004). Three indexes are chosen for full-resolution evaluation, i.e., spectral distortion index (D_λ), spatial distortion index (D_s), quality with no reference (QNR) (Alparone et al., 2008).

4.2 Reduced-resolution evaluation

Table 1 shows the objective performance of different methods on the reduced-resolution test set. It can be seen that the proposed CSN performs the best for all the reduced-resolution indexes, and this verifies the effectiveness of the proposed network. Comparing the last two rows of the table, it can be found that after adopting the semi-supervised training strategy, the performance on the test set declines slightly. Considering the difference between reduced-resolution data and full-resolution data, such a decline is acceptable, and the performance of CSN-SSL is still better than the comparison algorithms.

Fig. 5 shows the results of a reduced-resolution image. The results of MTF-GLP, GLP-SEGM, PNN and MSDCNN suffer from spectral distortion, and the color of the bare land is different from the reference image. The results of MMP and PNN suffer from blurring effects, and boundary of the buildings are unclear. In the enlarged viewers, MSDCNN, CSN and CSN-SSL well reconstruct the details of the roof, but in the result of the MSDCNN, obvious artifacts can be found. Only the proposed CSN and CSN-SSL generate pan-sharpened image similar to the reference image.

Method	CC	ERGAS	RMSE	SAM	SSIM
BDSD	0.9623	1.6443	36.1926	2.1291	0.8623
MMP	0.9473	2.2039	46.6171	2.3949	0.7798
MTF-GLP	0.9545	1.8849	39.9547	2.1689	0.8322
GLP-SEGM	0.9545	1.8849	39.9547	2.1689	0.8322
PNN	0.9423	2.1040	44.0697	2.2496	0.8070
MSDCNN	0.9645	1.5634	33.2124	2.0828	0.8721
CSN	**0.9762**	**1.2705**	**27.5365**	**1.5919**	**0.9062**
CSN-SSL	0.9747	1.3065	28.3795	1.6789	0.9023

Table 1. Objective performance of the pan-sharpening methods on reduced-resolution test set.

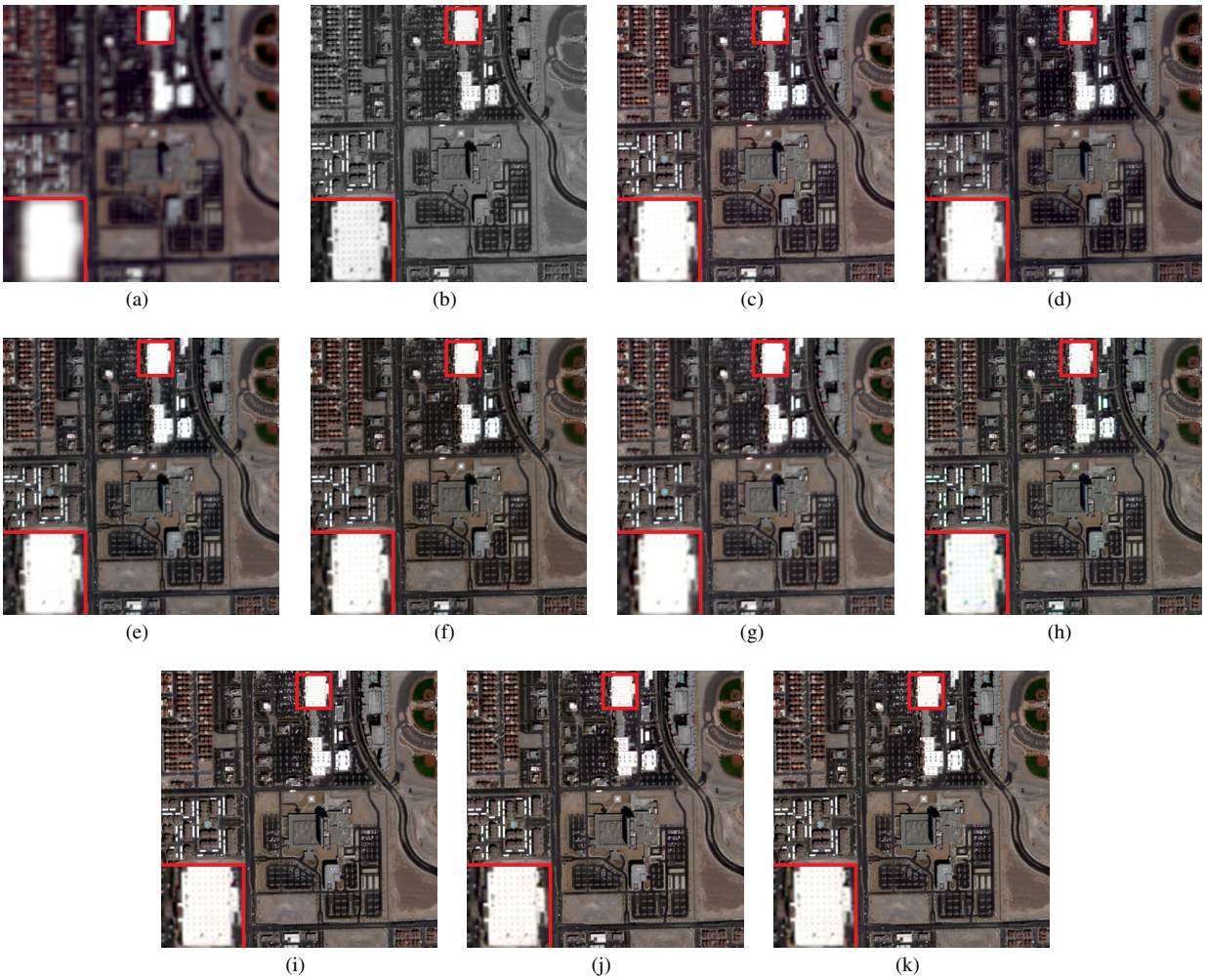


Figure 5. Comparison of pan-sharpening results obtained by different methods (reduced-resolution). (a) Low-resolution MS image. (b) PAN image. (c-j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, CSN, and CSN-SSL. (k) Reference image.

4.3 Full-resolution evaluation

Table 2 shows the objective performance of different methods on the full-resolution test set. Comparing the first four rows with the five to seven rows of the table, it can be seen that the deep learning based approaches can not show an advantage over the traditional approaches. But by adopting the semi-supervised training, the performance on the test set has been improved obviously and the proposed CSN-SSL achieves the best QNR index. This verifies the effectiveness of the proposed semi-supervised learning strategy.

Fig. 6 shows the results of a full-resolution image. In the en-

Method	Ds	Dl	QNR
BDSD	**0.0403**	0.0675	0.8953
MMP	0.0613	0.0709	0.8723
MTF-GLP	0.0455	0.0733	0.8848
GLP-SEGM	0.0432	0.0750	0.8853
PNN	0.0884	**0.0551**	0.8617
MSDCNN	0.0432	0.0723	0.8876
CSN	0.0547	0.0611	0.8877
CSN-SSL	0.0419	0.0580	**0.9027**

Table 2. Objective performance of the pan-sharpening methods on full-resolution test set.

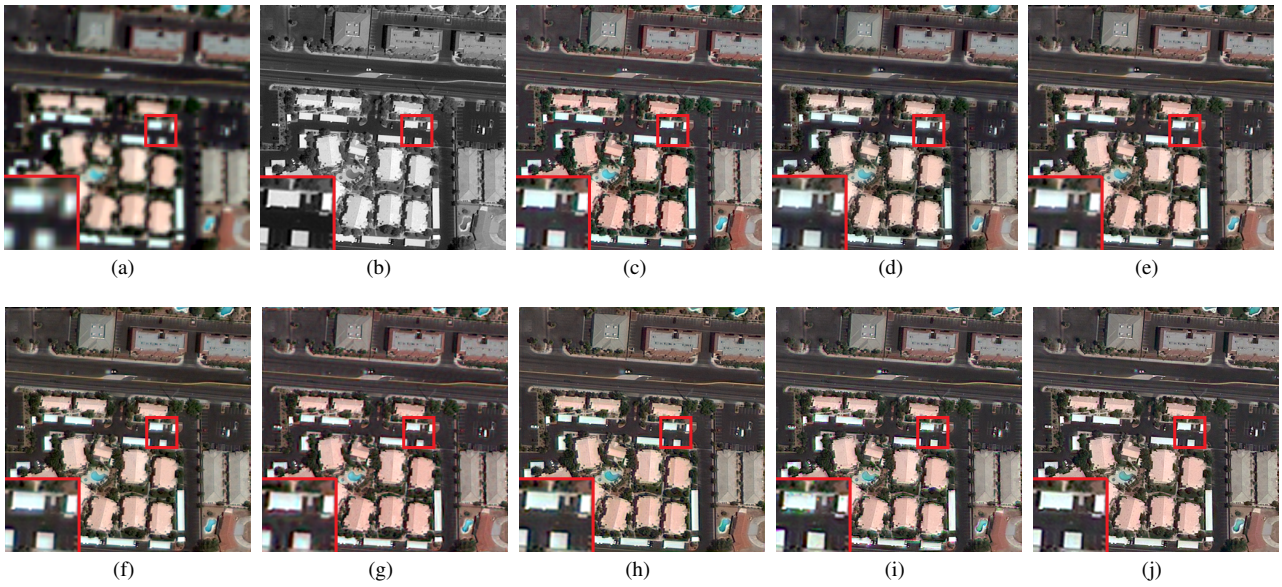


Figure 6. Comparison of pan-sharpening results obtained by different methods (full-resolution). (a) Low-resolution MS image. (b) PAN image. (c-j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, CSN, and CSN-SSL.

larged viewers, the results of BDSD, MTF-GLP, PNN, MSDCNN, and CSN suffer from severe artifacts. In the results of MMP and MTF-GLP, there are halos around the white buildings. Only the CSN-SSL generates image with good quality. Comparing the results of CSN and CSN-SSL, it can be found that the semi-supervised learning effectively suppresses the generation of artifacts.

5. CONCLUSION

In this paper, we have proposed a Component Substitution Network (CSN) for pan-sharpening. The CSN simulates the traditional component substitution approaches and overcomes their drawbacks. By adopting a semi-supervised learning strategy, the performance on full-resolution data is further improved. Compared with six existing state-of-the-art pan-sharpening methods, the results on reduced-resolution data and full-resolution data verify the proposed method has achieved state-of-the-art performance.

ACKNOWLEDGEMENTS

This work was supported in part by National Key Research and Development Program of China, Grant No. 2018YFB0505003, and National Natural Science Foundation of China with project number 41322010.

REFERENCES

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., 2002. Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 40(10), 2300–2312.

Alparone, L., Aiazzi, B., Baronti, S., Garzelli, A., Nencini, F., Selva, M., 2008. Multispectral and panchromatic data fusion assessment without reference. *Photogrammetric Engineering & Remote Sensing*, 74(2), 193–200.

Blum, A., Lafferty, J., Rwebangira, M. R., Reddy, R., 2004. Semi-supervised learning using randomized mincuts. *Proceedings of the twenty-first international conference on Machine learning*, 13.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4), 834–848.

Cheng, Y., 2019. Semi-supervised learning for neural machine translation. *Joint Training for Neural Machine Translation*, Springer, 25–40.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database. *2009 IEEE conference on computer vision and pattern recognition*, Ieee, 248–255.

Du, Q., Younan, N. H., King, R., Shah, V. P., 2007. On the performance evaluation of pan-sharpening techniques. *IEEE Geoscience and Remote Sensing Letters*, 4(4), 518–522.

Fasbender, D., Radoux, J., Bogaert, P., 2008. Bayesian data fusion for adaptable image pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 46(6), 1847–1857.

Garzelli, A., Nencini, F., Capobianco, L., 2008. Optimal MMSE pan sharpening of very high resolution multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, 46(1), 228–236.

Ghahremani, M., Ghassemian, H., 2016. A compressed-sensing-based pan-sharpening method for spectral distortion reduction. *IEEE Transactions on Geoscience and Remote Sensing*, 54(4), 2194–2206.

He, L., Rao, Y., Li, J., Chanussot, J., Plaza, A., Zhu, J., Li, B., 2019. Pansharpening via detail injection based convolutional neural networks. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(4), 1188–1204.

Huang, W., Xiao, L., Wei, Z., Liu, H., Tang, S., 2015. A new pan-sharpening method with deep neural networks. *IEEE Geoscience and Remote Sensing Letters*, 12(5), 1037–1041.

Kang, X., Li, S., Benediktsson, J. A., 2014. Pansharpening with matting model. *IEEE Transactions on Geoscience and Remote Sensing*, 52(8), 5088–5099.

Kuznietsov, Y., Stuckler, J., Leibe, B., 2017. Semi-supervised deep learning for monocular depth map prediction. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6647–6655.

Kwarteng, P., Chavez, A., 1989. Extracting spectral contrast in Landsat Thematic Mapper image data using selective principal component analysis. *Photogrammetric Engineering & Remote Sensing*, 55, 339–348.

Laben, C. A., Brower, B. V., 2000. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening. US Patent 6,011,875.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, K., Xie, W., Du, Q., Li, Y., 2019. DDLPS: Detail-Based Deep Laplacian Pansharpening for Hyperspectral Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(10), 8011–8025.

Lin, G., Milan, A., Shen, C., Reid, I., 2017. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1925–1934.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft coco: Common objects in context. *European conference on computer vision*, Springer, 740–755.

Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pan-sharpening by convolutional neural networks. *Remote Sensing*, 8(7), 594.

Mohammadzadeh, A., Tavakoli, A., Valadan Zoej, M. J., 2006. Road extraction based on fuzzy logic and mathematical morphology from pan-sharpened ikonos images. *The Photogrammetric Record*, 21(113), 44–60.

Nah, S., Hyun Kim, T., Mu Lee, K., 2017. Deep multi-scale convolutional neural network for dynamic scene deblurring. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3883–3891.

Nencini, F., Garzelli, A., Baronti, S., Alparone, L., 2007. Remote sensing image fusion using the curvelet transform. *Information Fusion*, 8(2), 143–156.

Nunez, J., Otazu, X., Fors, O., Prades, A., Pala, V., Arbiol, R., 1999. Multiresolution-based image fusion with additive wavelet decomposition. *IEEE Transactions on Geoscience and Remote sensing*, 37(3), 1204–1211.

Palsson, F., Sveinsson, J. R., Ulfarsson, M. O., 2014. A new pansharpening algorithm based on total variation. *IEEE Geoscience and Remote Sensing Letters*, 11(1), 318–322.

Ranzato, M., Szummer, M., 2008. Semi-supervised learning of compact document representations with deep networks. *Proceedings of the 25th international conference on Machine learning*, 792–799.

Restaino, R., Dalla Mura, M., Vivone, G., Chanussot, J., 2017. Context-adaptive pansharpening based on image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2), 753–766.

Rosenberg, C., Hebert, M., Schneiderman, H., 2005. Semi-supervised self-training of object detection models. *WACV/MOTION*, 2.

Scarpa, G., Vitale, S., Cozzolino, D., 2018. Target-adaptive CNN-based pansharpening. *IEEE Transactions on Geoscience and Remote Sensing*, 56(9), 5443–5457.

Souly, N., Spampinato, C., Shah, M., 2017. Semi and weakly supervised semantic segmentation using generative adversarial network. *arXiv preprint arXiv:1703.09695*.

Tu, T.-M., Huang, P. S., Hung, C.-L., Chang, C.-P., 2004. A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery. *IEEE Geoscience and Remote Sensing Letters*, 1(4), 309–312.

Wald, L., 2002. *Data fusion: definitions and architectures: fusion of images of different spatial resolutions*. Presses des MINES.

Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P. et al., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612.

Wei, Y., Yuan, Q., 2017. Deep residual learning for remote sensed imagery pansharpening. *2017 International Workshop on Remote Sensing with Intelligent Processing (RSIP)*, IEEE, 1–4.

Wei, Y., Yuan, Q., Shen, H., Zhang, L., 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. *IEEE Geoscience and Remote Sensing Letters*, 14(10), 1795–1799.

Weston, J., Ratle, F., Mobahi, H., Collobert, R., 2012. Deep learning via semi-supervised embedding. *Neural networks: Tricks of the trade*, Springer, 639–655.

Yuan, Q., Wei, Y., Meng, X., Shen, H., Zhang, L., 2018. 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*, 11(3), 978–989.

Yuhas, R. H., Goetz, A. F., Boardman, J. W., 1992. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm. *Summaries 4th JPL Airborne Earth Sci. Workshop*, 147–149.

Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019. Pan-Sharpening Using an Efficient Bidirectional Pyramid Network. *IEEE Transactions on Geoscience and Remote Sensing*, 57(8), 5549–5563.

Zhu, X., Ghahramani, Z., Lafferty, J. D., 2003. Semi-supervised learning using gaussian fields and harmonic functions. *Proceedings of the 20th International conference on Machine learning (ICML-03)*, 912–919.