Band-Independent Encoder–Decoder Network for Pan-Sharpening of Remote Sensing Images

Chi Liu, Yongjun Zhang, Shugen Wang, Mingwei Sun, Yangjun Ou, Yi Wan, and Xiu Liu

***Abstract*— Pan-sharpening is a fundamental task for remote sensing image processing. It aims at creating a high-resolution multispectral (HRMS) image from a multispectral (MS) image and a panchromatic (PAN) image. In this article, a new band-independent encoder–decoder network is proposed for pan-sharpening. The network takes a single band of the MS (BMS) image, the PAN image, and the low-resolution PAN (LRPAN) image as inputs. The output of the network is the corresponding band of high-resolution MS (HRBMS) image. In this way, the network can process MS images with any number of bands. The overall structure of the network consists of two encoder–decoder modules at low-resolution and high-resolution, respectively. An auxiliary LRPAN image is used to speed up the training and improve the performance. The partly shared network and hierarchical structure for low-resolution and high-resolution enable a better fusion of features extracted from different scales. With a fast fine-tuning strategy, the trained model can be applied to images from different sensors. Experiments performed on different data sets demonstrate that the proposed method outperforms several state-of-the-art pan-sharpening methods in both visual appearance and objective indexes, and the single-band evaluation results further verify the superiority of the proposed method.**

***Index Terms*— Band-independent, deep learning, encoder–decoder, pan-sharpening.**

I. Introduction

PAN-SHARPENING is a typical application of image fusion in the remote sensing field. It fuses low spatial resolution multispectral (MS) images and high spatial resolution panchromatic (PAN) images to construct high spatial resolution MS (HRMS) images. As a preprocessing for applications like image classification [1], object detection [2], and change detection [3], pan-sharpening provides synthetic images with both high spectral and spatial quality. With the emergence of newly launched satellite sensors, the problem becomes more complicated for the diverseness in spectral characteristics and spatial resolutions. Various algorithms have been proposed to address this problem.

Traditional approaches are component-substitution (CS)-based methods and multiresolution analysis (MRA)-based methods. The representative CS algorithms are principal component analysis (PCA) [4], intensity hue saturation (IHS) transform [5], and Gram–Schmidt (GS) sharpening [6]. Due to the spectral differences between the MS image and the PAN image, CS methods often encounter problems with spectral preservation and suffer from spectral distortions. Lately proposed CS methods concentrate on improving the spectral quality, strategies like adaptive coefficient calculation [7], partial replacement [8], local coefficient calculation [9] are used to reduce the spectral distortions, but the trade-off between spectral preservation and details injection remains a problem. The ideas of MRA methods are more straightforward than the CS methods. Details are extracted from the PAN image and then injected into the upsampled MS image. To obtain multiscale details, multiresolution analyses such as Laplacian pyramid [10], wavelet transform [11], curvelets transform [12], and non-subsampled contourlets transform [13] are used. As the quality of the output is sensitive to the details injected, insufficient details injection leads to blurring effects and excessive details injection results in artifacts and spectral distortions.

The third series of approaches is the model-based optimization (MBO) approaches. The main ideas of these approaches are to build models according to the relationship among the MS image, the PAN image, and the desired HRMS image. *A priori* constraints or different regularizations are formulated to solve the ill-posed inverse problem. Representative algorithms are sparsity regularization [14]–[17], Bayesian posterior probability [18], [19], variational models [20]–[24], and Markov random fields [25]. MBO methods are highly dependent on the regularization terms, sometimes the solution is unstable, and the time complexity of the MBO methods is much higher than many other algorithms.

There are also approaches based on geostatistics theory [26]–[28]. These approaches assume that when downsampling the pan-sharpened HRMS image to the MS resolution, the result should be identical to the MS image. This assumption makes the geostatistics-based approaches have a significant advantage in preserving the spectral characteristics

Manuscript received September 18, 2019; revised January 7, 2020 and February 11, 2020; accepted February 17, 2020. Date of publication February 26, 2020; date of current version June 24, 2020. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB0505003, in part by the National Natural Science Foundation of China under Grant 41871368 and Grant 41801386, in part by the National High-Score Major Special Projects under Grant 50-H31D01-0508-13/15, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20180797. *(Corresponding author: Yongjun Zhang.)*

Chi Liu, Yongjun Zhang, Shugen Wang, Mingwei Sun, Yangjun Ou, and Yi Wan are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: liuchi@whu.edu.cn; zhangyj@whu.edu.cn; wangsg@whu.edu.cn; mingweis@whu.edu.cn; ouyangjun@whu.edu.cn; yi.wan@whu.edu.cn).

Xiu Liu is with the Beijing Institute of Space Mechanics and Electricity, Beijing 100094, China (e-mail: liuxiu0725@163.com).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2020.2975230

of the MS image. However, the regression models used by geostatistics-based approaches are linear, and it is difficult for the models to formulate the complex pan-sharpening process when the spectral range of the MS bands is not fully covered by that of the PAN image.

Recently, deep learning has been introduced to the field of remote sensing image processing tasks, several pan-sharpening networks [29]–[40] have been designed, and their performance has shown great potential. The high nonlinearity of the convolutional neural network makes it practical to deal with the pan-sharpening problem. However, most of the existing networks are not customized for the pan-sharpening task. The common cases are adopting networks for other computer vision tasks like image super-resolution [41], [42] or semantic segmentation [43]–[46], the characteristics of the MS image and the PAN image are ignored. What is more, the networks can only process MS images composed of specified bands. Different models have to be trained for images with a different number of bands.

In this article, probing into the problems in pan-sharpening, we propose a band-independent encoder–decoder network for pan-sharpening. The main contributions of this article lie in: 1) inputs of the network are a single band of the MS (BMS) image and the PAN image, which makes the network robust to spectral difference and permits the network to handle images with any number of bands; 2) an auxiliary low-resolution PAN (LRPAN) image is used to speed up the training and to improve the performance; 3) the partly shared network and hierarchical structure enable a better fusion of features from different scales; and 4) single-band evaluation and comparison of the pan-sharpening results.

The remainder of this article is organized as follows. Section II briefly introduces the background knowledge of the encoder–decoder network and the existing deep learning-based pan-sharpening methods. 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, discussion and conclusion are given in Section V.

II. Related Work

A. Encoder–Decoder Network

The encoder–decoder network was first proposed by Hinton and Salakhutdinov [47] in 2006 for data reduction. High-dimensional data contain a lot of redundant information and noise. By training a multilayer "encoder" network, it can be converted to low-dimensional codes that only keep the most critical information, and a similar multilayer "decoder" network can recover the data from the codes. With a well-trained encoder and decoder network, the conversion between high-dimensional data and low-dimensional data is accomplished.

With the rise of deep learning, the encoder–decoder networks have been successfully applied to many applications. In the field of image processing, the encoder–decoder networks are associated with the image resolution and have a broader meaning. Typically, an encoder–decoder network contains an encoder module that gradually reduces the feature maps and captures higher semantic information and a decoder module that gradually recovers the low-resolution encoded feature maps to full input resolution feature maps. The encoder–decoder networks have been widely used in computer vision tasks including human pose estimation [48], object detection [49], [50], semantic segmentation [43]–[46], and single image super-resolution [41], [42], [51]–[53].

An extensively researched problem with the encoder–decoder network in image processing task is the preservation of details. The boundary information lost in the encoder module is difficult to be recovered in the decoder module. To overcome this problem, many efforts have been made. In [50], each decoder upsampled input feature maps and added them to the corresponding encoded feature maps to produce the input of the next decoder. In [54], the locations of the maximum feature value in each pooling window was memorized for each encoder feature map, and the decoder network upsampled its input feature maps using the memorized max-pooling indices from the corresponding encoded feature maps. In [55], a U-shaped structure network (U-Net) was proposed, the upsampled feature maps in the decoder module were concatenated with the corresponding encoded feature maps. The U-Net is simple, yet effective, and it has been adopted and improved by many researchers [56]–[59]. Our proposed network is also based on the idea of U-Net.

B. Deep Learning-Based Pan-Sharpening

The first time deep learning used for pan-sharpening was in [29], modified sparse denoising auto-encoder (MSDA) network was trained using the low-resolution and high-resolution PAN image patches, and then used to predict HRMS images from LRMS images. In [30], a similar network was proposed. In [31], a deep metric learning method was proposed to learn a refined geometric multimanifold neighbor embedding via multiple nonlinear deep neural networks, and by the assumption that MS patches and PAN patches formed the same geometric manifolds in two distinct spaces, the high-resolution MS image patches were estimated. These methods transform the pan-sharpening task into image superresolution, and only the PAN images are involved in the training. The spectral characteristics of the LRMS image are ignored, and the outputs of the networks often encounter problems with spectral quality.

Other methods take the MS image and the PAN image together as the inputs of the network. In the preprocessing step, these methods upsample the MS image and concatenate it with the PAN image to compose a synthetic image. Masi *et al.* [32] designed a shallow network with only three convolutional layers for pan-sharpening. Scarpa *et al.* [33] improved architectures of [32], they used a deeper network with residual-learning and added a target-adaptive tuning phase, after an efficient fine-tuning, significant performance could be achieved even for across sensor images. Wei *et al.* [34] used a 11-layer deep residual network for pan-sharpening, and they adopted convolution kernels of larger size

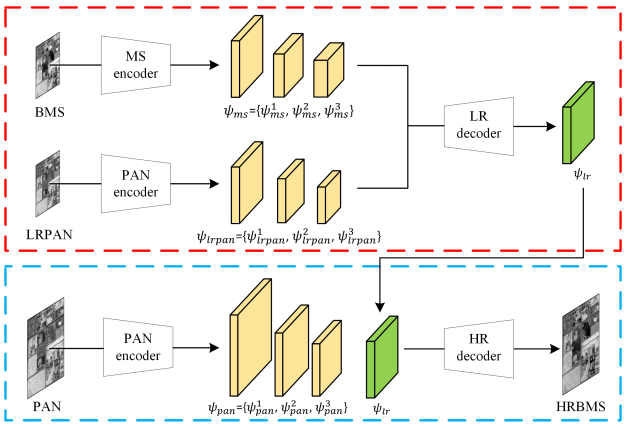


Fig. 1. Structure of the proposed network. The low-resolution encoder–decoder module is inside the red dotted area, and the high-resolution encoder–decoder module is in the blue dotted area.

for better performance. In [35], a two-stream network was proposed, convolutional kernels of different sizes were used to extract features of different scales. Yao *et al.* [36] built a pan-sharpening model employing the structure of U-Net. However, these methods process the MS image and the PAN image in the same way and ignore their individual characteristics. In addition, the upsampling operation in preprocessing step increases the quantity of computation.

In [37], a bidirectional pyramid structure was proposed to process the MS image and the PAN image separately following the general idea of MRA. Multilevel details were extracted from the PAN image and injected into the MS image to reconstruct the pan-sharpened image. Similarly, in [38], a details injection network was proposed for pan-sharpening. However, in both the networks, two streams of the networks are connected only using an adding operation, which makes it difficult for the networks to explore the relationship between the MS image and the PAN image. Yang *et al.* [39] trained network parameters in the high-pass filtering domain rather than the image domain to preserve the spatial structure. Shao and Cai [40] extracted feature maps from the MS image and the PAN image separately and then concatenated them to reconstruct the details. Nevertheless, these two networks have not considered multiscale analysis, so the adaptability of the networks remains to be explored.

Another common drawback of existing networks for pan-sharpening is that the number of bands for the MS image has to be predefined. The input of the networks is related to the number of bands for the MS image, so are the network structures (mainly the dimension of convolutional kernels). Different models have to be trained for MS images with a different number of bands, and the applicability of the models is limited.

III. Methodology

In this section, we will describe the design methodology of the proposed network. Fig. 1 shows the overall structure of the proposed network. The network consists of a low-resolution encoder–decoder module and a high-resolution encoder–decoder module. The low-resolution encoder–decoder module copes with the BMS image and the LRPAN image to obtain a low-resolution feature map. The high-resolution encoder–decoder module takes the PAN image and the output of the low-resolution module as inputs to reconstruct the band of high-resolution MS (HRBMS).

A. Single-Band Inputs

To the best of our knowledge, all the existing pan-sharpening networks take the multiband MS image and the PAN image as inputs. Although the networks can output the multiband pan-sharpened image end to end, they have a limited range of applications. First, the networks are sensitive to spectral difference, and the model trained on the image composed of certain bands is difficult to be fine-tuned for images composed of other bands. Another limitation of existing networks is that the number of bands for the MS image is fixed when the bands of the input are not matched with the model, the network cannot work. Therefore, different models have to be trained for images with a different number of bands.

For most satellite sensors, the spectral range of the MS bands are nonoverlapped or little overlapped, making the relationship among the MS bands ambiguity. Usually, this relationship is connected with the PAN image based on the assumption that the PAN image can be simulated by a linear combination of the MS bands. However, the complexity of ground objects and different spectral responses of sensors make this relationship unreliable and unstable. Thus, many traditional approaches [10]–[12] process the MS image band by band.

Considering the limitations of multiband inputs, we propose our band-independent network. The inputs of the network are single-band images, and the output of the network is the corresponding HRBMS image. Multiscale details and spectral information are extracted from the PAN image and the BMS image, respectively, to reconstruct the HRBMS image. In the training, each BMS image is used as a training sample, so fewer training image pairs are needed. Different bands of the MS image share the same network, making the network robust to spectral characteristics of the inputs and work for images from more satellite sensors. Moreover, the band-independent property permits the trained model to predict MS images with any number of bands.

B. Auxiliary LRPAN Image

The resolution of the PAN image is four times that of the MS image, making it difficult for the network to process them together. So in most of the existing networks, the MS image is upsampled to the resolution of the PAN image so that they can be processed synchronously. However, the upsampling operation does not improve the resolution of the MS image essentially, rudely connecting the MS image with the PAN image in this way not only brings in artifacts but also increases computational cost.

When adopting the encoder–decoder framework for pan-sharpening, to get feature maps of the same scale, the network

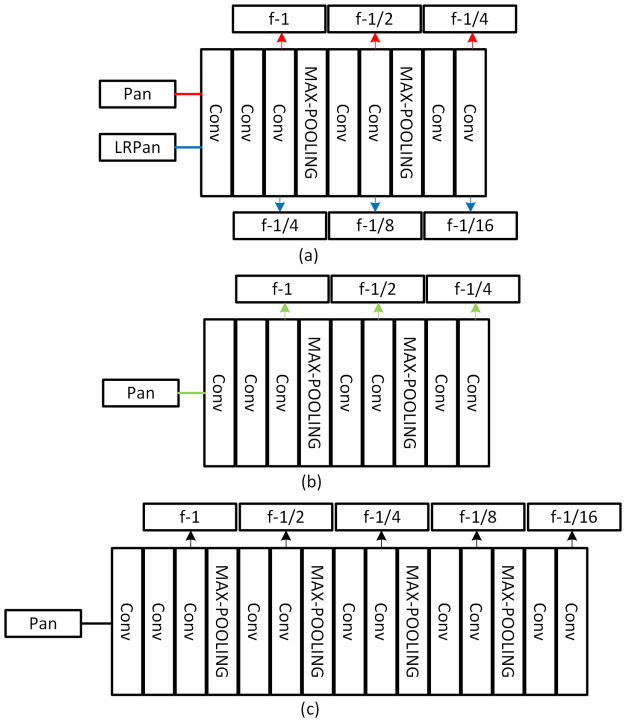


Fig. 2. Different encoders for the PAN/LRPAN image. (a) Proposed encoder. (b) Shares the same structure with (a) but only takes the PAN image as input. (c) Takes the PAN image as input and outputs similar features with (a).

for the PAN image has to be deeper than the network for the MS image, which leads to an increase in network depth. Moreover, when feature maps of the PAN image and the MS image come from different depths of the network, it becomes difficult for the network to converge. To avoid these problems, in the proposed network, we add an LRPAN as input to reduce the depth of the network and speed up the training. The same encoder network is used by the PAN image and the LRPAN image, which not only reduces the size of our pan-sharpening network but also makes the network able to extract multiscale details and effective for satellite images of different spatial resolutions. In addition, the LRPAN image can serve as the bridge between the MS image and the PAN image to help the network converge quickly.

A comparison between networks with and without the LRPAN image is shown in Fig. 2. Fig. 2(a) is the structure of the proposed encoder for PAN/LRPAN image, by taking the PAN image and the LRPAN image together as inputs, with a shallow network, feature maps at five different scales are extracted. The network in Fig. 2(b) shares the same network with Fig. 2(a), but only takes the PAN image as input, and only extracts feature maps at three scales. The network in Fig. 2(c) takes the PAN as input and extracts feature maps at five scales; however, it is much deeper than the proposed network. The performance comparison between the networks will be discussed in Section IV in detail.

C. Hierarchical Network

To take full use of the LRPAN image, a hierarchical structure is adopted. As shown in Fig. 1, the proposed network consists of a low-resolution module and a high-resolution module. In the low-resolution module, the MS image and the LRPAN image are encoded separately and then decoded together into a fusion feature map at the resolution of the MS image. In the high-resolution module, the PAN image is encoded into multiscale feature maps and then decoded with the fusion feature map from the low-resolution module. Since the LRPAN image is of the same resolution with the MS image, it helps the network in the low-resolution module to better preserve the spectral characteristics of the MS image. The participant of the LRPAN image is also beneficial for the fusion of the low-resolution feature maps and the encoded PAN feature maps.

To get the multilevel expression of the input MS image, we adopt the first six layers of the VGG16 [60] as an encoder network for the MS image. Some modifications have been made for our specific task. Since the input of the MS encoder network is a single-band image, the kernel dimension of the first convolutional layer is changed to 1. The residual learning [61] solves the gradient vanishing problem and learns the residual between the input and output, which is especially suitable for the pan-sharpening task, so we add skip connections [61] to all the convolutional layers. The outputs of the MS encoder network are feature maps at different scales whose number of channels are 64, 128, and 256, respectively. The encoder network for the PAN image and the LRPAN image is similar to that of the MS image. Since the feature maps of the PAN image and the LRPAN image concentrate on image details which degrade dramatically after downsampling, there is no need to double the number of channels for the feature maps after each downsampling layer, so the number of channels for the output feature maps is fixed to 64 in the PAN/LRPAN encoder network. Using a fixed number of channels also largely reduces the memory cost of the network.

The encoded MS feature maps are concatenated with the LRPAN feature maps, and then two convolutional layers are used to fuse the feature maps, after which an upsampling layer is used to upsample the feature maps. Like the U-Net, each time the feature maps are upsampled, they are concatenated with the encoded feature maps at the same scale. The output of the low-resolution decoder is a 64-channel feature map which is of the same resolution with the input MS image. This together with the encoded PAN feature maps is used as the inputs of the high-resolution decoder. The main structure of the high-resolution decoder is similar to the low-resolution one. The output of the high-resolution decoder is a single-band image, i.e., the desired HRBMS.

D. Objective Function

The proposed network is an image generative network [62], which aims at producing an image as similar to the target image as possible. There are many objective functions to choose from, such as content loss [62], perceptual loss [63], and adversarial loss [64]. Among the objective functions, the perceptual loss and adversarial loss are complicated to be calculated, additional networks are introduced for loss calculation. So, these two loss functions are more suitable

Algorithm 1 Network Training

Input: training samples, training epochs
Output: trained model

Initialization: Set learning rate, training epochs;
for *n = 1, ..., nEpochs* **do**
  for *m = 1, ..., nSamples* **do**
    Get multiscale MS feature maps $\boldsymbol{\Psi_{ms}} = \{\boldsymbol{\psi^1_{ms}}, \boldsymbol{\psi^2_{ms}}, \boldsymbol{\psi^3_{ms}}\}$ from the BMS image using the MS encoder;
    Get multiscale LRPAN feature maps $\Psi_{lrpan} = \{\psi^1_{lrpan}, \psi^2_{lrpan}, \psi^3_{lrpan}\}$ from the LRPAN image using the PAN encoder;
    Decode Ψ_{ms} and Ψ_{lrpan} into low-resolution feature map Ψ_{lr} using the low-resolution decoder;
    Get multiscale PAN feature maps $\Psi_{pan} = \{\psi^1_{pan}, \psi^2_{pan}, \psi^3_{pan}\}$ from the PAN image using the PAN encoder;
    Decode Ψ_{lr} and Ψ_{pan} using the high-resolution decoder to get the HRBMS;
    Calculate the loss between the output HRBMS and the ground truth;
    Back propagation and update the network parameters;

for high-level computer vision task but inefficient for pan-sharpening. The content loss is more basic and straightforward in calculation, and the common ones are ℓ_1-loss and ℓ_2-loss. ℓ_2-loss is sensitive to significant errors and tends to produce smooth images, so we use the ℓ_1-loss, which is efficient and edge-sensitive as our objective function

$$\text{loss} = \frac{1}{w * h} \sum_{i=1}^{w} \sum_{j=1}^{h} \| \boldsymbol{I}(i, j) - \boldsymbol{G}(i, j) \|_1 \tag{1}$$

where i and j are the pixel indexes, w and h are the width and height of the image, I is the output HRBMS of the network, G is the ground truth (the desired HRBMS image), and $\| \cdot \|_1$ is the absolute value function. For each BMS, a loss is calculated and used to update network parameters. Training is carried out by optimizing the objective function using Adam optimization algorithm [65] based on back propagation [66]. The training process of the proposed network is summarized in Algorithm 1.

IV. Experiments

A. Experimental Setup

In the proposed network, each BMS image can be used as a training sample, so much fewer image patches are needed. The training set is composed of 1000 GeoEye-1 (GE1) image patches from six GE1 images, 800 GaoFen-2 (GF2) image patches from four GF2 images and 400 WorldView (WV) image patches from three WV images. The validation set and test set are both composed of 200 GE1 image patches, 200 GF2 image patches, and 50 WV image patches. It should be noted that since our network takes the BMS as input, the actual number of samples for the network is equal to the product of the number of image patches and the number of bands in the MS image. Following Wald's protocol, spatially degraded images are used as inputs, and the original MS images are used as the reference images. For full-resolution evaluation, other 150 GE1 image patches, 200 GF2 image patches, and 80 WV image patches are prepared. The patch sizes of the MS image are 100 × 100 pixels for the degraded experiments and 200 × 200 pixels for the full-resolution experiments.

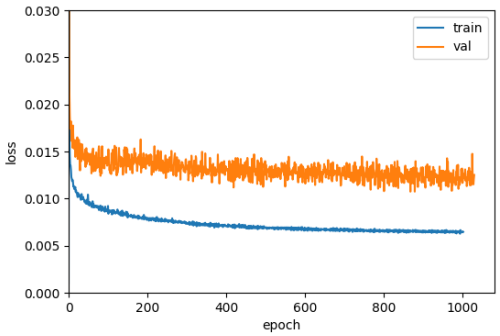


Fig. 3. Loss curves of the GE1 training set and the validation set.

Besides our proposed method, seven state-of-the-art pan-sharpening methods are used for comparison, i.e., BDSD [67], MMP [68], MTF-GLP [10], GLP-SEGM [69], PNN [33], MSDCNN [35], and BDPN [37]. Among them, BDSD is a well-known CS-based method; MMP is a creative approach which is a combination of the CS method and the MBO method; MTF-GLP and GLP-SEGM are MRA-based methods which achieve state-of-the-art performance; the other three methods (i.e., PNN, MSDCNN, and BDPN) are deep learning-based approaches mentioned in Section II-B. To make a comprehensive assessment, test results of different methods are evaluated by a series of indexes, both at reduced-resolution and full-resolution. Six indexes are chosen for reduced-resolution evaluation, i.e., correlation coefficient (CC), relative dimensionless global error in synthesis (ERGAS) [70], root-mean-square error (RMSE), spectral angle mapper (SAM) [71], universal image quality index (Q2n) [72], and structural similarity index (SSIM) [73]. Three indexes are chosen for full-resolution evaluation, i.e., spectral distortion index (D_λ), spatial distortion index (D_s), quality with no reference index (QNR) [74].

B. Training Details

Since training a model with all the training data took more time, we trained the model using only the GE1 training set, and then fine-tuned the model on the GF2 training set and WV training set for GF2 image and WV image, respectively. We selected the GE1 training set because it has the largest number of training samples, and the resolution of GE1 image (0.5 m for the PAN image) falls in between the resolution of GF2 image (1.0 m for the PAN image) and WV image (0.5/0.31 for the WV2/WV3 PAN image). Impressive results have been achieved in this way, and the detailed results are given in Section IV-D.

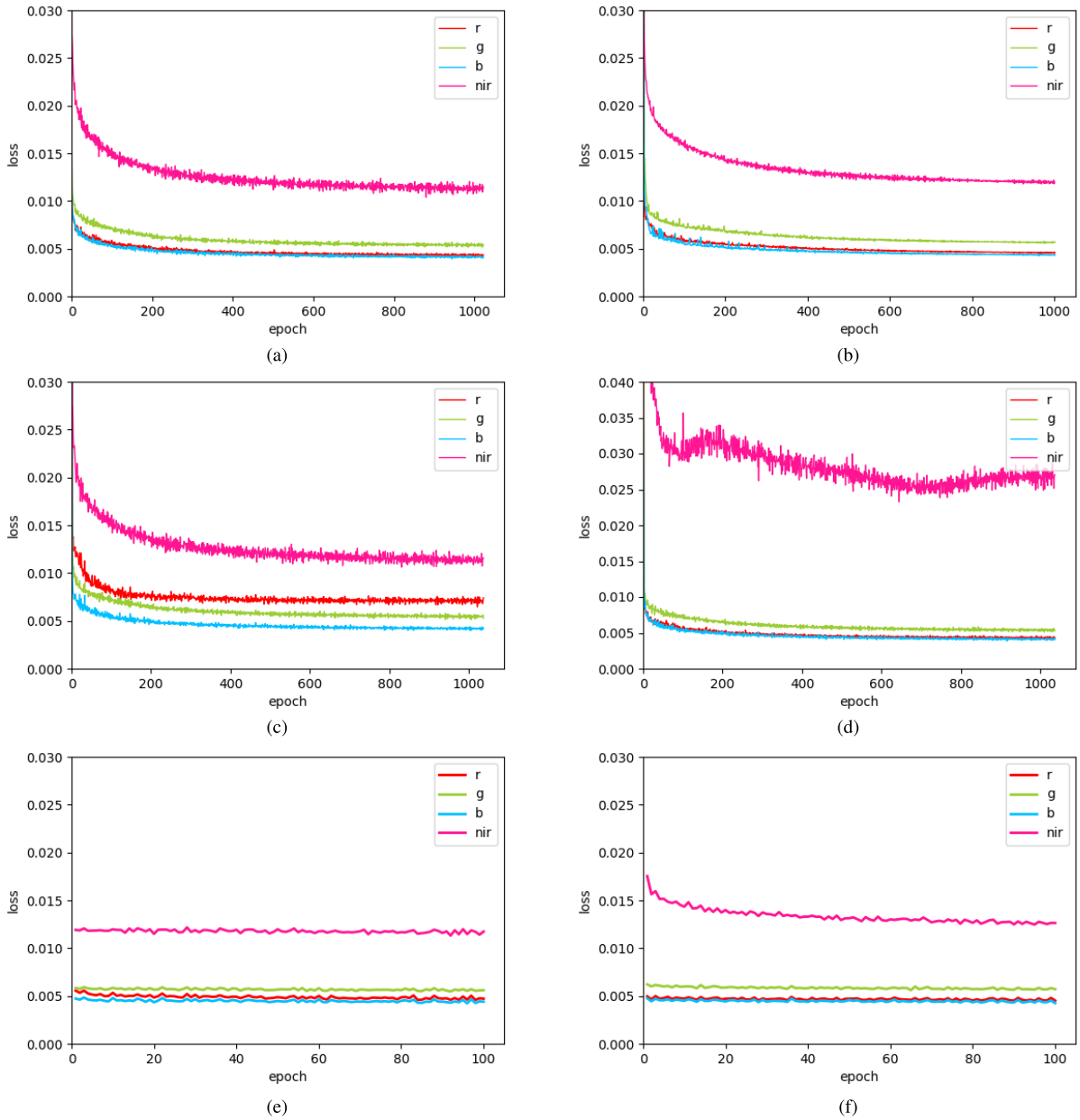


Fig. 4. Loss curves of each band using different training sets. The training sets of (a)–(d) are images from all bands, images from a single band, images from b/g/nir band and images from r/g/b band, respectively. (e) and (f) Loss curves when fine-tuning the model of (c) and (d) on images from all bands, respectively.

In training, the initial learning rate was set to 1.0×10^{-4}; the learning rate descent factor was set to 0.8 every 100 epochs. The model converged after about 600 epochs, which took about 6 hours. In the fine-tuning stage, the learning rate was set to 1.0×10^{-5}, the model was fine-tuned on the GF2 training set and WV training set for 5 epochs, respectively, which took about 5 min. The batch size for training and fine-tuning were 12. The training patches were randomly clipped to 64×64 and 256×256 pixels for MS images and PAN images, respectively, before being put into the network.

The loss curves of the training set and the validation set are shown in Fig. 3. Since the validation set is a composition of images from different sensors, its loss curve is not as stable as the training set. However, the overall trend of the loss curves is similar. It means that even the model is trained only on GE1 image patches, and its performance on image patches from other sensors improve. This verifies the robustness of the proposed network.

C. Ablation Study

To demonstrate the band-independent property of the network, we grouped the images by the band and used different band groups as training sets. First, we trained the network with images from all bands, at the same time, the loss curves of every single band were recorded, and the result is shown

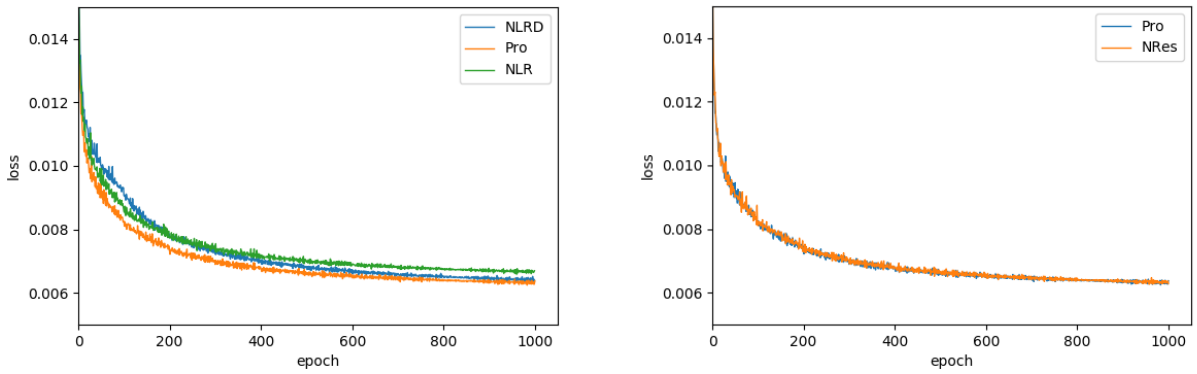


Fig. 5. Loss curves of networks with different structures. (a) Plots of the loss curves of networks with different encoders for the PAN/LRPAN image. (b) Plots of the loss curves of the networks with and without skip connections.

in Fig. 4(a). Though the losses of each band converge at different values, all bands decline synchronously and converge at about 600 epoch. This demonstrates the band-independent property of the network to some degree. However, the losses of each band are different, especially the loss of the near-infrared band is much heavier than the other three bands. It is doubtful whether the network has achieved its best performance on each band. So, we added another experiment which used images from a single band as the training set and trained the network. For each of the R, G, B, and NIR bands, a model was trained and the loss curve was recorded. The four loss curves are plotted in Fig. 4(b). It can be seen that the loss curves are almost the same as that in Fig. 4(a), which demonstrates that images from different bands are independent and can be trained together.

Also, we trained the network with images from three bands and used images from the remaining band as a validation data set. Specifically, two models were trained, one was trained on images from the green, blue, and near-infrared bands, and the other one was trained on images from the red, green, and blue bands. The loss curves are shown in Fig. 4(c) and (d), respectively. It can be seen that losses of the bands used for training are similar to that in Fig. 4(b), which further verifies the band-independent property. However, for the bands not involved in the training (the validation band), the losses are not as well as that in Fig. 4(b). The loss curve of the red band in Fig. 4(c) converges at the 200 epoch, and its final loss is heavier than that in Fig. 4(b), which means the network does not achieve its best performance on the red band. The loss curve of the near-infrared band in Fig. 4(d) does not converge, and the loss is much heavier than that in Fig. 4(b). The poor performance on the validation data sets is mainly caused by the spectral characteristics of different bands. To further explore the network's ability to learn spectral characteristics, we fine-tuned the model with images from all bands, and the loss curves are shown in Fig. 4(e) and (f). In both the figures, the losses of the three bands which have been trained hardly change; but for the bands not involved in the training, the loss curves drop rapidly and converge to a value similar to that in Fig. 4(a) after about 50 epochs. This demonstrates the good adaptability of the network, and that is why the trained model can be used to process images from other sensors with a brief fine-tuning.

To verify the effectiveness of the LRPAN image, we trained networks without the LRPAN image and compared them with the proposed network. To make a fair comparison, the structures of the networks are similar except for the encoder for the PAN image. The compared networks ED_NLR and ED_NLRD are obtained by replacing the encoder for the PAN/LRPAN image with the encoders introduce in Fig. 2(b) and (c), respectively. Since the PAN encoder in ED_NLR outputs only three scale features, features of the PAN image did not participate in the low-resolution module in ED_NLR. The loss curves of the networks on GE1 training set are shown in Fig. 5(a), it can be seen that the proposed network converges most quickly and achieves the minimum loss. The ED_NLR also converges quickly; however, its final loss is heavier than the proposed network, and this verifies that the LRPAN image helps to improve the performance of the network. We think the reason is that it is difficult for the low-resolution module to extract useful information without the LRPAN image. The final loss of the ED_NLRD is close to the proposed network; however, it converges much slower than the other two networks, and this is mainly caused by its deep network structure.

Another experiment we do is to explore the impact of the skip connections. We removed all the skip connections in the network to obtain a comparison network which we called ED_NRes. Since the skip connections do not add any parameters to the network, the ED_NRes has exactly the same number of parameters with the proposed network. The loss curve of the ED_NRes on GE1 training set is shown in Fig. 5(b). For ease of comparison, the loss curve of the proposed network is also plotted. It can be seen that the two loss curves are almost overlapped, which shows that the skip connections have not helped much in the proposed network. We think the reason might be that the skip connection is effective for vanishing gradients problem in deep networks, but the proposed network is relatively shallow. Since the skip connection is a common technique, and it does not cause a negative impact on the network, we still add skip connections to our network.

TABLE I
OBJECTIVE PERFORMANCE OF THE PAN-SHARPENING METHODS ON GE1 TEST SET

Methods	reduced resolution indexes						full resolution indexes		
	ERGAS	CC	SAM	RMSE	Q4	SSIM	D_λ	D_s	QNR
BDSD	1.5962	0.9641	2.0998	35.3728	0.9099	0.8639	0.0916	0.0519	0.8616
MMP	2.1396	0.9495	2.3579	45.5277	0.8329	0.7831	0.0858	0.0644	0.8554
MTF-GLP	1.8272	0.9566	2.1380	39.0322	0.8808	0.8355	0.0947	0.0474	0.8626
GLP-SEGM	1.7154	0.9597	2.1478	37.5992	0.8978	0.8549	0.0991	0.0478	0.8580
PNN	2.7800	0.9363	2.2469	51.3390	0.8480	0.8076	0.0873	0.0953	0.8259
MSDCNN	2.2676	0.9401	2.4592	46.3583	0.7644	0.7921	**0.0705**	0.0633	0.8708
BDPN	1.5201	0.9687	2.0052	33.2281	0.9216	0.8906	0.1164	**0.0357**	0.8520
Proposed	**1.1274**	**0.9817**	**1.5189**	**24.3535**	**0.9457**	**0.9228**	0.0818	0.0375	**0.8840**

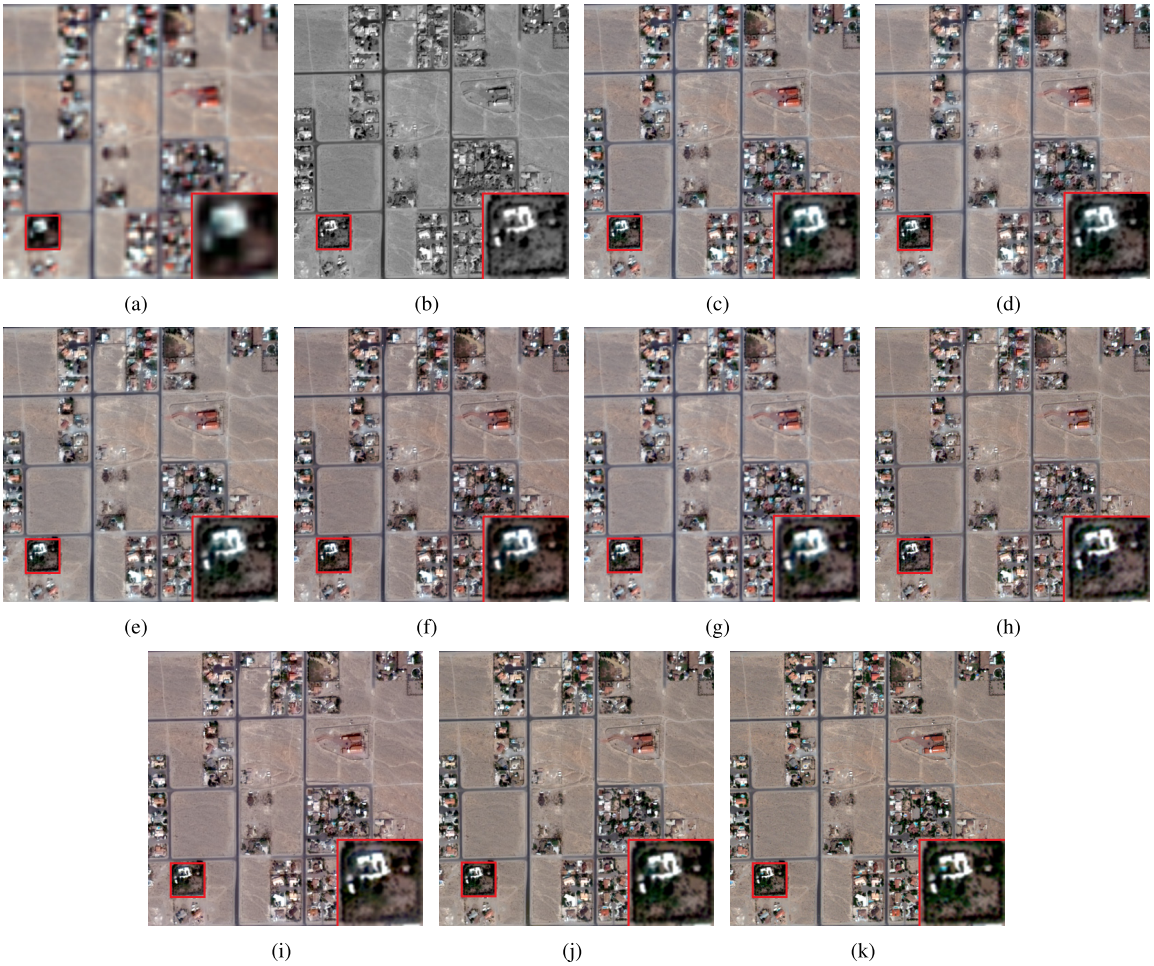


Fig. 6. Comparison of pan-sharpening results obtained by different methods (downsampled GE1 image). (a) Low-resolution MS image. (b) PAN image. (c)–(j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, BDPN, and the proposed method. (k) Reference image.

D. Evaluation on the Test Set

In this section, we compare the performance of different methods on the test sets. Objective indexes are listed in tables, with the best result for each index shown in boldface. For visual comparison, sample image patches are displayed at reduced resolution and full resolution. All the images are rendered by ArcGIS Desktop [75] with default parameters; for the MS images, the red, green, and blue bands (band 3, 2, 1 for GE1 and GF2, band 5, 3, 2 for WV) are chosen for display.

Table I shows the objective performance of different methods on the GE1 test set. It can be seen that the proposed method performs the best for all the reduced-resolution indexes, as well as the QNR index, whereas MSDCNN and BDPN get the best D_λ index and D_s index, respectively. Fig. 6 shows the results of a GE1 test image patch at reduced-resolution. The results of BDSD, MMP MTF-GLP suffer from color distortions, the color of bare lands is different from that in the reference image. In the enlarged views, the color of the vegetation regions is abnormal in the results of GLP-SEGM, MSDCNN, and BDPN, indicating that these

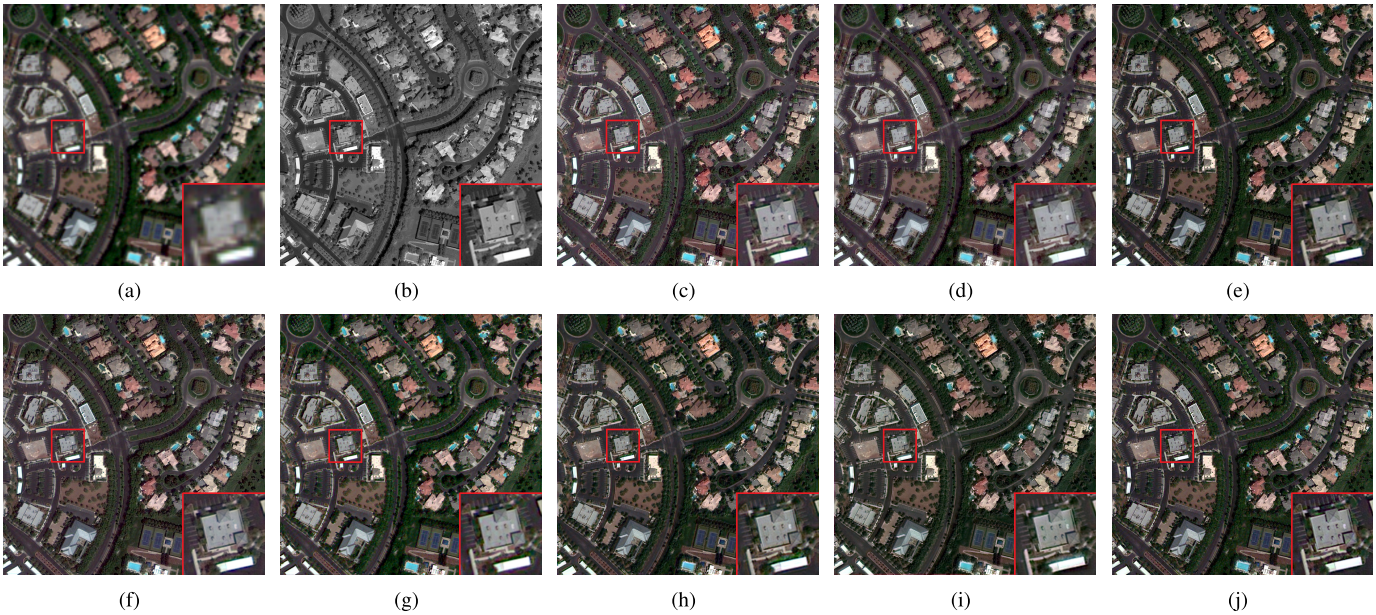


Fig. 7. Comparison of pan-sharpening results obtained by different methods (GE1 image). (a) Low-resolution MS image. (b) PAN image. (c)–(j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, BDPN, and the proposed method.

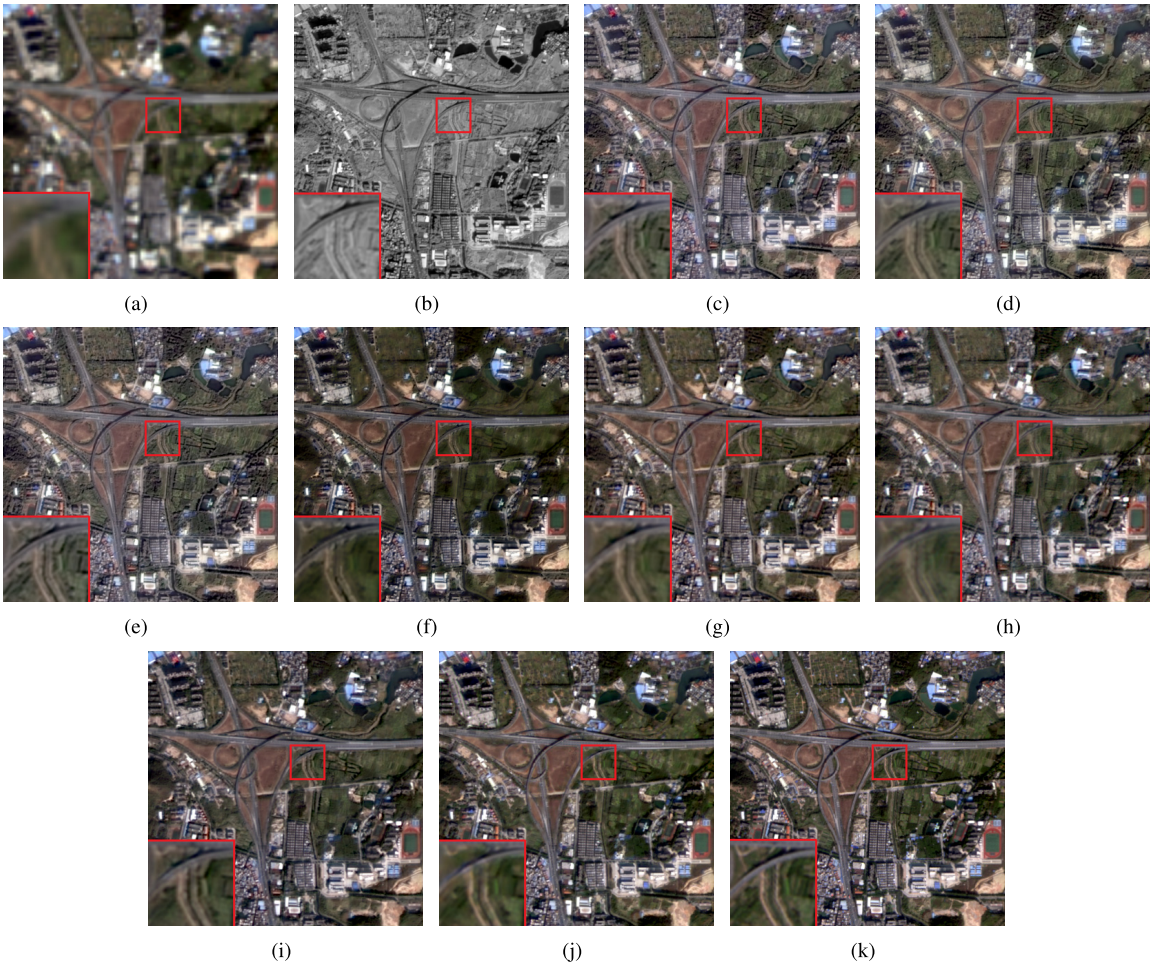


Fig. 8. Comparison of pan-sharpening results obtained by different methods (downsampled GF2 image). (a) Low-resolution MS image. (b) PAN image. (c)–(j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, BDPN, and the proposed method. (k) Reference image.

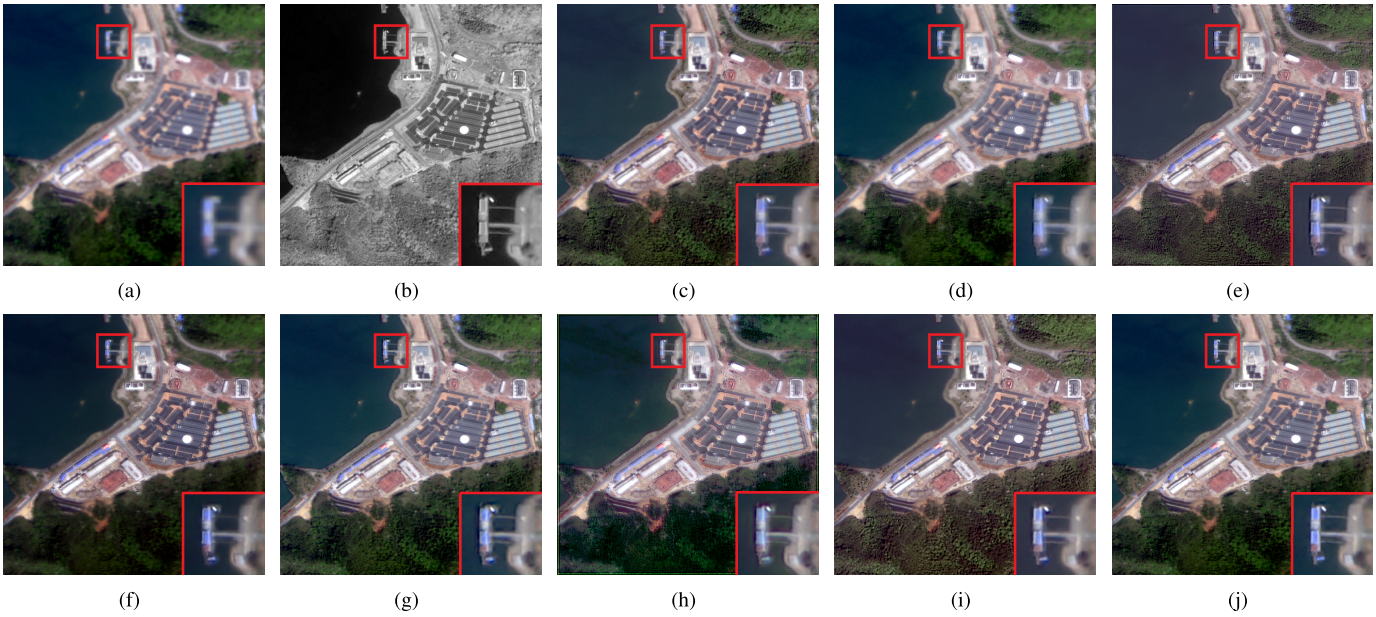


Fig. 9. Comparison of pan-sharpening results obtained by different methods (GF2 image). (a) Low-resolution MS image. (b) PAN image. (c)–(j) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, BDPN, and the proposed method.

TABLE II
OBJECTIVE PERFORMANCE OF THE PAN-SHARPENING METHODS ON GF2 TEST SET

Methods	reduced resolution indexes						full resolution indexes		
	ERGAS	CC	SAM	RMSE	Q4	SSIM	D_λ	D_s	QNR
BDSD	1.7539	0.9342	1.6538	27.3710	0.8408	0.7493	0.1231	0.0419	0.8414
MMP	1.7669	0.9305	1.7184	27.5758	0.8093	0.7219	0.0834	0.0934	0.8298
MTF-GLP	1.7503	0.9303	**1.6267**	27.3891	0.8264	0.7351	0.1856	0.0641	0.7654
GLP-SEGM	1.7901	0.9288	1.7664	27.5890	0.8268	0.7329	0.1386	0.0551	0.8149
PNN	1.7976	0.9300	1.6926	27.8131	0.8415	0.7714	0.0975	0.0669	0.8421
MSDCNN	1.9783	0.9135	2.0209	30.1217	0.7741	0.6819	**0.0345**	0.1250	0.8447
BDPN	1.7252	0.9295	1.7062	27.0345	0.8433	0.7573	0.1342	**0.0378**	0.8330
Proposed	**1.6687**	**0.9396**	1.7373	**25.5874**	**0.8693**	**0.7795**	0.0996	0.0606	**0.8455**

results also suffer from color distortions. As for spatial quality, only the proposed method successfully recovers the house in the enlarged view, the results of other methods suffer from a different level of blurring effects. Fig. 7 shows a full-resolution experiment performed on GE1. The results of BDSD, GLP-SEGM, and BDPN suffer from color distortions, and the color of vegetation regions in the results of these methods is different from that in the MS image. The result of MMP suffers from blurring effect, the road centerlines and the pedestrian crossing are not very clear. In the enlarged views, the result of the proposed method is more clear and natural, whereas in the results of other methods, there are different levels of halo effects around the white building. The result of PNN also suffers from serious artifacts.

Table II shows the results on the GF2 test set, and the proposed method achieves the best performance on reduced-resolution indexes except for SAM. As for the full-resolution indexes, in general, all the methods achieve good performance regarding the D_s index, whereas the D_λ index is not that satisfactory. The MMP method and the BDPN achieve the best D_λ and D_s, respectively, and the proposed method achieves the best QNR. Fig. 8 shows a GF2 experiment at reduced resolution. The results of MTF-GLP, GLP-SEGM, and BDPN suffer from spectral distortions, the color of the lawns are different from that of the ground truth. In the enlarged views, it can be seen that the proposed method achieves the image with the best quality, the color of the lawns are the most natural and the boundaries of the roads are clearest. Fig. 9 shows a GF2 experiment at full resolution. The results of MTF-GLP, GLP-SEGM, MSDCNN, and BDSD suffer from serious color distortions, the color of the vegetation regions and the water regions in these results are different from the MS image. The result of BDSD also suffers from a little color distortion, the vegetation region in it is brighter than that in the MS image. In the enlarged views, only the proposed method successfully reconstructs the building, and the results of other methods suffer from different levels of blurring effects around the boundary of the buildings.

Table III shows the results on the WV test set. Since the BDPN method mainly works on images with four-bands,

TABLE III
OBJECTIVE PERFORMANCE OF THE PAN-SHARPENING METHODS ON WV TEST SET

Methods	reduced resolution indexes						full resolution indexes		
	ERGAS	CC	SAM	RMSE	Q4	SSIM	D_λ	D_s	QNR
BDSD	6.5544	0.9097	8.0252	74.3980	0.7759	0.7060	0.0894	0.0882	0.8311
MMP	7.5052	0.8917	8.2561	84.8455	0.7028	0.6300	**0.0411**	**0.0382**	**0.9222**
MTF-GLP	6.8549	0.9007	7.7817	77.5482	0.7590	0.6815	0.0800	0.0536	0.8725
GLP-SEGM	6.7220	0.9052	7.5373	75.7281	0.7559	0.6838	0.0679	0.0488	0.8882
PNN	7.8496	0.8820	8.0474	87.7456	0.6667	0.6062	0.1306	0.0919	0.7909
MSDCNN	5.8918	**0.9320**	8.3940	67.1021	0.7691	0.7453	0.0655	0.0603	0.8807
Proposed	**5.5525**	0.9296	**6.8537**	**63.2163**	**0.7969**	**0.7522**	0.0822	0.0669	0.8563

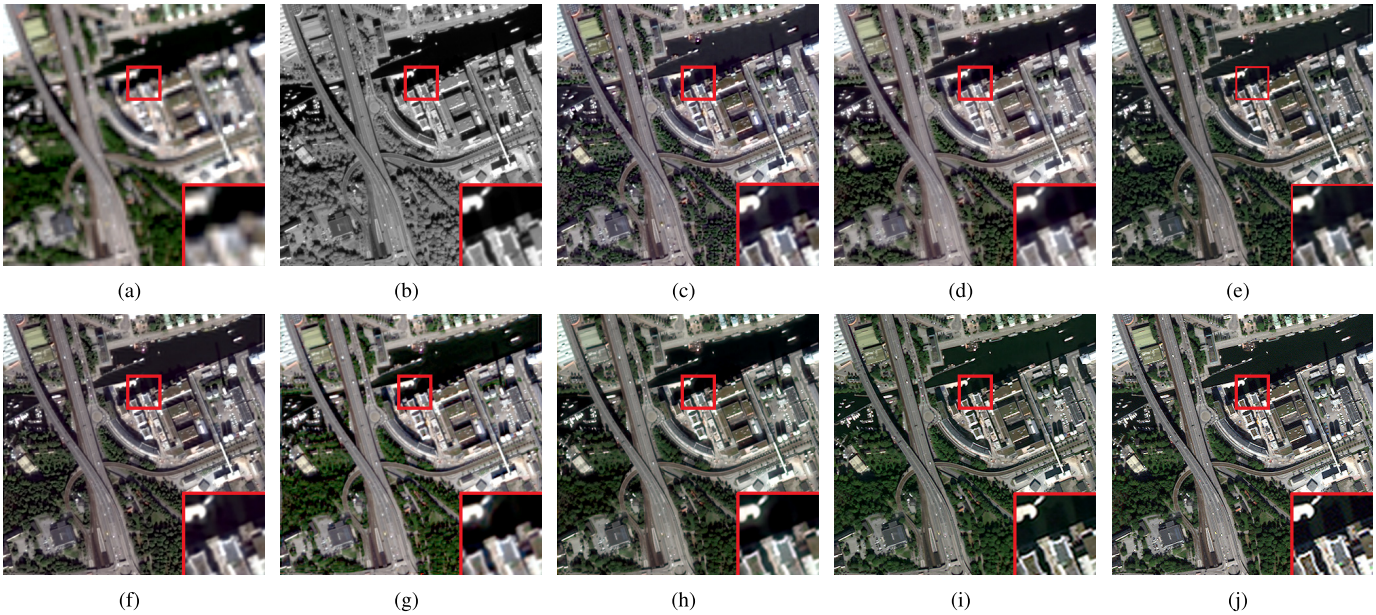


Fig. 10. Comparison of pan-sharpening results obtained by different methods (downsampled WV image). (a) Low-resolution MS image. (b) PAN image. (c)–(i) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, and the proposed method. (j) Reference image.

and there are not enough data to train an eight-band model, it does not participate in the comparison. From the table, it can be seen that the proposed method achieves the best performance on reduced-resolution indexes except for CC. As for full-resolution evaluation, MMP achieves the best performance, followed by GLP-SEGM. The deep learning-based methods (PNN, MSDCNN, and the proposed method) do not perform as good as the traditional ones (MMP, GLP-SEGM), the reason is that the deep learning models cannot properly deal with the noise in full-resolution WV test images. Fig. 10 gives an example of a reduced-resolution WV experiment. The result of BDSD suffers from spectral distortions as indicated by the abnormal color of the vegetation regions. The results of MMP, MTF-GLP, GLP-SEGM, PNN, and MSDCNN suffer from different levels of spatial distortions; in the enlarged views, the shadows of the buildings are blurring. Only the proposed method produces a clear pan-sharpened image similar to the reference image. Fig. 11 gives an example of the full-resolution WV experiment. To reduce the influence of noise, we selected a patch with little noise. It can be seen in the enlarged views that the results of BDSD and PNN suffer from severe artifacts around the ground objects. The results of MTF-GLP, GLP-SEGM, and MSDCNN are a little blurring, and there are halos around the highlighted objects. MMP and the proposed method achieve clear pan-sharpened images.

E. Single-Band Evaluation

Since our proposed method works on single-band images, we also compare the single-band performance of our method with the contrast algorithms. By comparing the result of different bands, some laws and regulations have been discovered. To the best of our knowledge, this is the first time the pan-sharpening results have been evaluated in single-band. Since most of the full-resolution indexes cannot be applied to a single-band image, the comparisons are made at reduced resolution. Specifically, four indexes, i.e., ERGAS, CC, Q, and SSIM, which can be used for single-band image, are selected as evaluation indexes.

The single-band evaluation results of the GE1 test set are shown in Fig. 12. By comparing the results of different methods, it can be seen that the proposed method robustly achieves the best performance for all the indexes of each

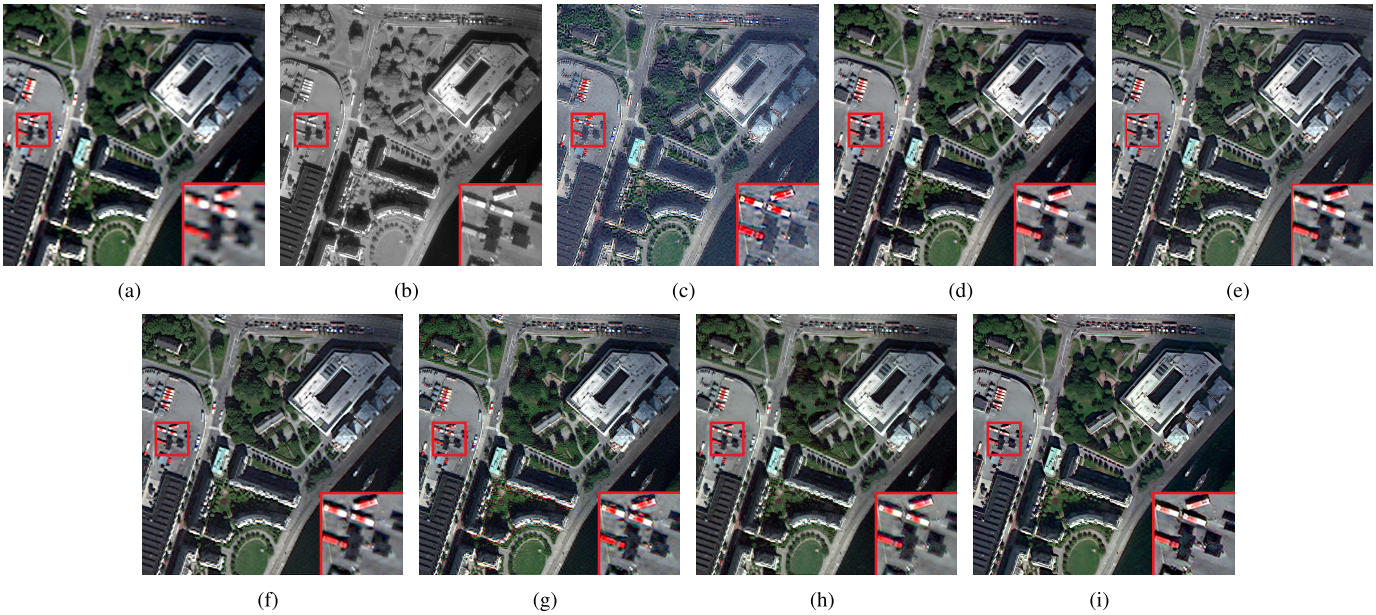


Fig. 11. Comparison of pan-sharpening results obtained by different methods (WV image). (a) Low-resolution MS image. (b) PAN image. (c)–(i) Pan-sharpening results of BDSD, MMP, MTF-GLP, GLP-SEGM, PNN, MSDCNN, and the proposed method.

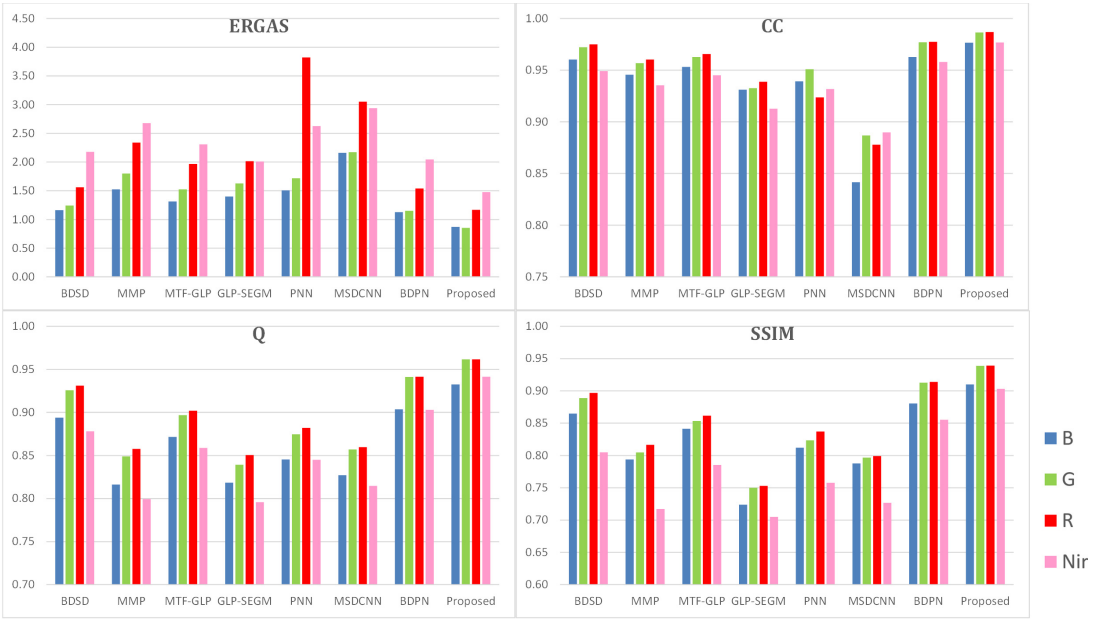


Fig. 12. Single-band evaluation of pan-sharpening results obtained by different methods (GE1).

band, which verifies the superiority of the proposed method. Besides, the proposed method performs more equally on each band compared to other methods. By comparing the results of different bands, we also find some laws. For most methods, the red band achieves the best performance regarding CC, Q, and SSIM, and the performance on the near-infrared band is not as good as the other three bands. It can be concluded that the red band is the most correlated with the PAN image, whereas the near-infrared band is the least. However, for the ERGAS index, the red band performs worse than the blue band and the green band. The reason is that the red band has a lower average value than the blue and green bands in the GE1 image.

The single-band evaluation results of the GF2 test set are shown in Fig. 13. Comparing the results of different methods, the proposed method almost achieves the best performance for all indexes, except that it achieves the second performance for the CC and SSIM on the red band. Comparing the results of different bands, similar to the GE1 image, the proposed method performs more equally on each band, and all methods perform the worst on the near-infrared band. The difference is that the compared methods perform the best on the red band, but the proposed method performs the best on the blue band, which shows the band-independent training in the proposed method better explores the relationship between single-band MS image and the PAN image and

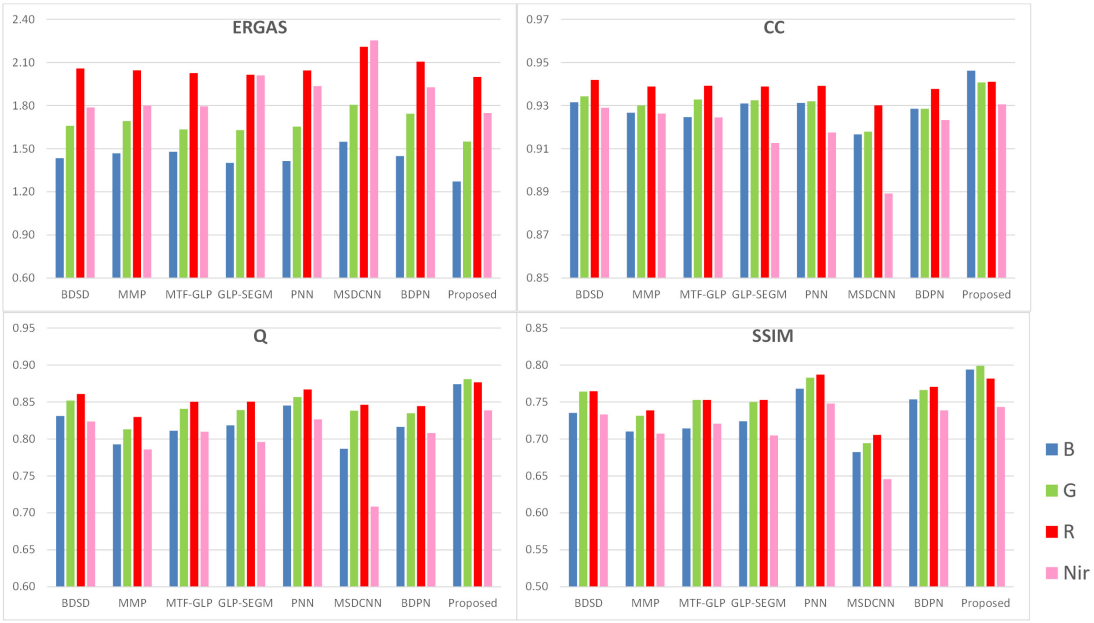


Fig. 13. Single-band evaluation of pan-sharpening results obtained by different methods (GF2).

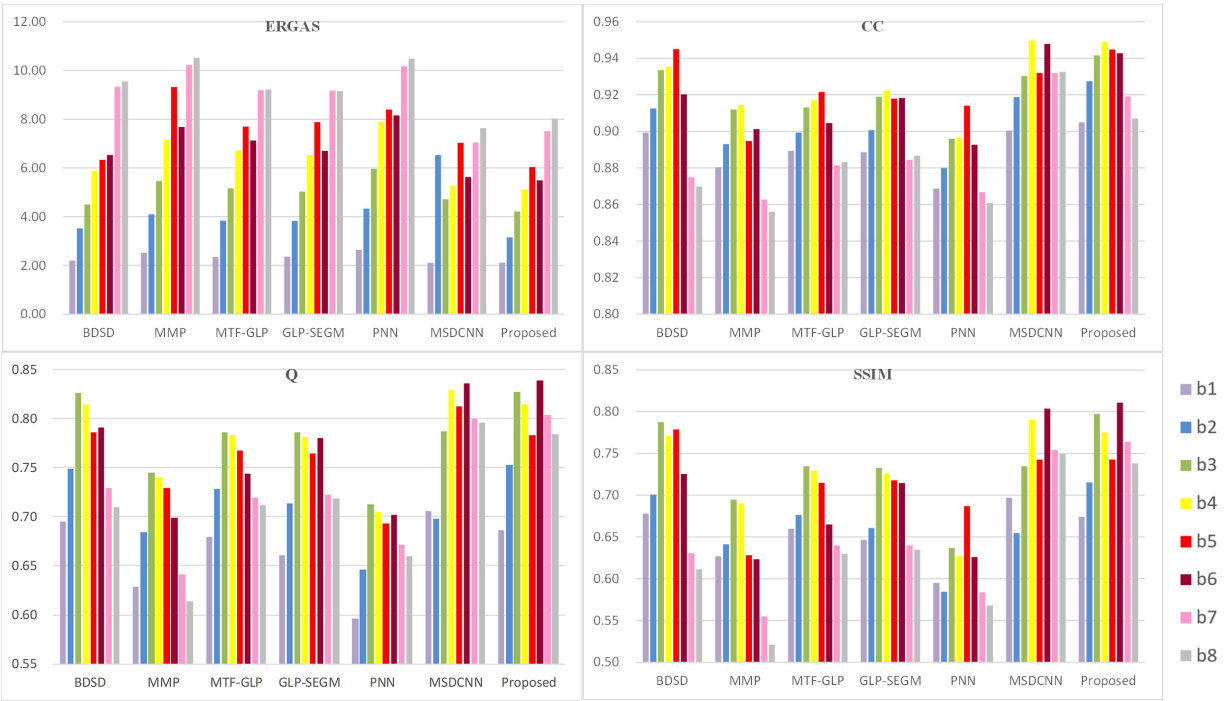


Fig. 14. Single-band evaluation of pan-sharpening results obtained by different methods (WV).

breaks through the limitations of traditional methods to some extent.

The single-band evaluation results of the WV test set are shown in Fig. 14. Like the previous process, we first compare the performance of different methods. Since the WV image has more bands compared to GE1 and GF2, the comparison is more challenging. However, it can be seen that the proposed method still achieves the best performance on most indexes for single bands. In particular, a dramatic improvement has been made on band 7 and band 8 compared to most of the contrast algorithms, and this demonstrates once again the stability of the proposed method. Then, we compare the performance upon different bands. Generally, all methods perform better on band 3 to band 6, but different methods achieve the best performance on different bands. For example, the proposed method and MSDCNN perform the best on band 6, while BDSD, MMP, MTF-GLP, and GLP-SEGM perform the best on band 3, PNN performs the best on band 5. The reason for better

performance on these bands is that the spectral ranges of these bands are covered by the spectral range of the PAN image. However, it is difficult to explain why all the methods perform not very well on the blue band (band 2), whose spectral range is also fully covered by the spectral range of the PAN image. The results of two near-infrared band (band 7 and band 8) are very similar, because the spectral range of the two bands are partly overlapped, and the spectral responses of ground objects to these two bands have strong consistency.

V. Conclusion

In this article, we have proposed a new band-independent encoder–decoder network for pan-sharpening. By adopting single-band input and a reused encoder module, the network is robust to spectral characteristics and spatial resolution of the input images. Much fewer samples are needed to train the model, and with a fast fine-tuning strategy, the trained model can be applied to images from different sensors. Compared with seven existing state-of-the-art pan-sharpening methods, the results on different data sets verify the superiority and robustness of the proposed method. However, similar to the BDPN, the proposed network can only be used to process MS and PAN images whose resolutions differ by four times, and with some modifications, the network can process MS and PAN images whose resolutions differ by 2^n times. We will explore to optimize the proposed method for MS and PAN images with any level of scaling factors.

References

[1] Q. Du, N. H. Younan, R. King, and V. P. Shah, "On the performance evaluation of pan-sharpening techniques," *IEEE Geosci. Remote Sens. Lett.*, vol. 4, no. 4, pp. 518–522, Oct. 2007.

[2] A. Mohammadzadeh, A. Tavakoli, and M. J. Valadan Zoej, "Road extraction based on fuzzy logic and mathematical morphology from pan-sharpened ikonos images," *Photogramm. Rec.*, vol. 21, no. 113, pp. 44–60, Mar. 2006.

[3] C. Souza, "Mapping forest degradation in the eastern Amazon from SPOT 4 through spectral mixture models," *Remote Sens. Environ.*, vol. 87, no. 4, pp. 494–506, Nov. 2003.

[4] P. S. Chavez, Jr., and A. Y. Kwarteng, "Extracting spectral contrast in landsat thematic mapper image data using selective principal component analysis," *Photogramm. Eng. Remote Sens.*, vol. 55, no. 3, pp. 339–348, 1989.

[5] T.-M. Tu, P. S. Huang, C.-L. Hung, and C.-P. Chang, "A fast intensity-hue-saturation fusion technique with spectral adjustment for IKONOS imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 1, no. 4, pp. 309–312, Oct. 2004.

[6] C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening," U.S. Patent 6 011 875, Jan. 4, 2000.

[7] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, and T. Wittman, "An adaptive IHS pan-sharpening method," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 4, pp. 746–750, Oct. 2010.

[8] J. Choi, K. Yu, and Y. Kim, "A new adaptive component-substitution-based satellite image fusion by using partial replacement," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 295–309, Jan. 2011.

[9] J. Liu, Y. Hui, and P. Zan, "Locally linear detail injection for pansharpening," *IEEE Access*, vol. 5, pp. 9728–9738, 2017.

[10] B. Aiazzi, L. Alparone, S. Baronti, and A. Garzelli, "Context-driven fusion of high spatial and spectral resolution images based on oversampled multiresolution analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 40, no. 10, pp. 2300–2312, Jan. 2002.

[11] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, and R. Arbiol, "Multiresolution-based image fusion with additive wavelet decomposition," *IEEE Trans. Geosci. Remote Sens.*, vol. 37, no. 3, pp. 1204–1211, May 1999.

[12] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, Apr. 2007.

[13] V. P. Shah, N. H. Younan, and R. L. King, "An efficient pan-sharpening method via a combined adaptive PCA approach and contourlets," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1323–1335, May 2008.

[14] S. Li and B. Yang, "A new pan-sharpening method using a compressed sensing technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 738–746, Feb. 2011.

[15] S. Li, H. Yin, and L. Fang, "Remote sensing image fusion via sparse representations over learned dictionaries," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 9, pp. 4779–4789, Sep. 2013.

[16] M. Ghahremani and H. Ghassemian, "A compressed-sensing-based pan-sharpening method for spectral distortion reduction," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2194–2206, Apr. 2016.

[17] R. Fei, J. Zhang, J. Liu, F. Du, P. Chang, and J. Hu, "Convolutional sparse representation of injected details for pansharpening," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1595–1599, Oct. 2019.

[18] D. Fasbender, J. Radoux, and P. Bogaert, "Bayesian data fusion for adaptable image pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1847–1857, Jun. 2008.

[19] H. Yin, "PAN-guided cross-resolution projection for local adaptive sparse representation-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4938–4950, Jul. 2019.

[20] F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "A new pansharpening algorithm based on total variation," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 318–322, Jan. 2014.

[21] 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, Sep. 2016.

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

[23] G. Vivone, P. Addesso, R. Restaino, M. Dalla Mura, and J. Chanussot, "Pansharpening based on deconvolution for multiband filter estimation," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 540–553, Jan. 2019.

[24] X. Fu, Z. Lin, Y. Huang, and X. Ding, "A variational pan-sharpening with local gradient constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, p. 10.

[25] M. Xu, H. Chen, and P. K. Varshney, "An image fusion approach based on Markov random fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 12, pp. 5116–5127, Dec. 2011.

[26] E. Pardo-Igúzquiza, M. Chica-Olmo, and P. M. Atkinson, "Downscaling cokriging for image sharpening," *Remote Sens. Environ.*, vol. 102, nos. 1–2, pp. 86–98, May 2006.

[27] Y. Tang, P. M. Atkinson, and J. Zhang, "Downscaling remotely sensed imagery using area-to-point cokriging and multiple-point geostatistical simulation," *ISPRS J. Photogram. Remote Sens.*, vol. 101, pp. 174–185, Mar. 2015.

[28] Q. Wang, W. Shi, and P. M. Atkinson, "Area-to-point regression kriging for pan-sharpening," *ISPRS J. Photogram. Remote Sens.*, vol. 114, pp. 151–165, Apr. 2016.

[29] W. Huang, L. Xiao, Z. Wei, H. Liu, and S. Tang, "A new pan-sharpening method with deep neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 5, pp. 1037–1041, May 2015.

[30] A. Azarang and H. Ghassemian, "A new pansharpening method using multi resolution analysis framework and deep neural networks," in *Proc. 3rd Int. Conf. Pattern Recognit. Image Anal. (IPRIA)*, Apr. 2017, pp. 1–6.

[31] Y. Xing, M. Wang, S. Yang, and L. Jiao, "Pan-sharpening via deep metric learning," *ISPRS J. Photogram. Remote Sens.*, vol. 145, pp. 165–183, Nov. 2018.

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

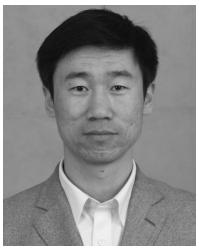
[33] G. Scarpa, S. Vitale, and D. Cozzolino, "Target-adaptive CNN-based pansharpening," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 9, pp. 5443–5457, Sep. 2018.

[34] Y. Wei, Q. Yuan, H. Shen, and L. Zhang, "Boosting the accuracy of multispectral image pansharpening by learning a deep residual network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1795–1799, Oct. 2017.

[35] Q. Yuan, Y. Wei, X. Meng, H. Shen, and L. Zhang, "A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening," *IEEE J. Sel.Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 3, pp. 978–989, Mar. 2018.
[36] W. Yao, Z. Zeng, C. Lian, and H. Tang, "Pixel-wise regression using U-Net and its application on pansharpening," *Neurocomputing*, vol. 312, pp. 364–371, Oct. 2018.
[37] Y. Zhang, C. Liu, M. Sun, and Y. Ou, "Pan-sharpening using an efficient bidirectional pyramid network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5549–5563, Aug. 2019.
[38] L. He *et al.*, "Pansharpening via detail injection based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 4, pp. 1188–1204, Apr. 2019.
[39] 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)*, Oct. 2017, pp. 5449–5457.
[40] Z. Shao and J. Cai, "Remote sensing image fusion with deep convolutional neural network," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1656–1669, May 2018.
[41] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2014, pp. 184–199.
[42] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1646–1654.
[43] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.
[44] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1925–1934.
[45] J. Fu, J. Liu, Y. Wang, J. Zhou, C. Wang, and H. Lu, "Stacked deconvolutional network for semantic segmentation," *IEEE Trans. Image Process.*, to be published.
[46] W. Zhao, S. Du, Q. Wang, and W. J. Emery, "Contextually guided very-high-resolution imagery classification with semantic segments," *ISPRS J. Photogram. Remote Sens.*, vol. 132, pp. 48–60, Oct. 2017.
[47] G. E. Hinton, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[48] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 483–499.
[49] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[50] A. Shrivastava, R. Sukthankar, J. Malik, and A. Gupta, "Beyond skip connections: Top-down modulation for object detection," 2016, *arXiv:1612.06851*. [Online]. Available: http://arxiv.org/abs/1612.06851
[51] L. Liebel and M. Körner, "Single-image super resolution for multispectral remote sensing data using convolutional neural networks," *ISPRS-Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 41, pp. 883–890, Jun. 2016.
[52] J. M. Haut, R. Fernandez-Beltran, M. E. Paoletti, J. Plaza, A. Plaza, and F. Pla, "A new deep generative network for unsupervised remote sensing single-image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6792–6810, Nov. 2018.
[53] S. Pendurkar, B. Banerjee, S. Saha, and F. Bovolo, "Single image super-resolution for optical satellite scenes using deep deconvolutional network," in *Proc. Int. Conf. Image Anal. Process.* Berlin, Germany: Springer, 2019, pp. 410–420.
[54] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
[55] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2015, pp. 234–241.
[56] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Berlin, Germany: Springer, 2016, pp. 424–432.
[57] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 5, pp. 749–753, May 2018.
[58] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Berlin, Germany: Springer, 2018, pp. 3–11.
[59] T. Falk *et al.*, "U-Net: Deep learning for cell counting, detection, and morphometry," *Nature Methods*, vol. 16, no. 1, pp. 67–70, Dec. 2018.
[60] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[62] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.
[63] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2016, pp. 694–711.
[64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.
[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[66] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[67] 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, Jan. 2008.
[68] X. Kang, S. Li, and J. A. Benediktsson, "Pansharpening with matting model," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 8, pp. 5088–5099, Aug. 2014.
[69] R. Restaino, M. Dalla Mura, G. Vivone, and J. Chanussot, "Context-adaptive pansharpening based on image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 753–766, Feb. 2017.
[70] L. Wald, *Data Fusion: Definitions Architectures: Fusion Images Different Spatial Resolutions*. Paris, France: Presses des MINES, 2002.
[71] R. H. Yuhas, A. F. H. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 4th JPL Airborne Earth Sci. Workshop*, 1992, pp. 147–149.
[72] A. Garzelli and F. Nencini, "Hypercomplex quality assessment of Multi/Hyperspectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 662–665, Oct. 2009.
[73] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[74] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, and M. Selva, "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.*, vol. 74, no. 2, pp. 193–200, Feb. 2008.
[75] *Arcgis Desktop: Release 10*, R. ESRI, Environmental Systems Research Institute, Redlands, CA, USA, 2011.

Chi Liu was born in 1994. He received the B.S. degree in photogrammetry and remote sensing from Wuhan University (WHU), Wuhan, China, in 2015, where he is currently pursuing the Ph.D. degree in photogrammetry and remote sensing.

He is engaged in high-spatial resolution remote sensing image processing, including pan-sharpening, color balancing for remote sensing imagery, and change detection.



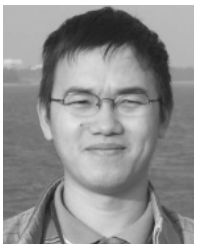
Yongjun Zhang was born in 1975. He received the B.S., M.S., and Ph.D. degrees from Wuhan University (WHU), Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, WHU. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, integration of LiDAR point clouds and images, and 3-D city reconstruction.

Dr. Zhang is the winner of the Second-Class National Science and Technology Progress Award in 2017. He has been supported by the New Century Excellent Talents in University from the Ministry of Education of China in 2007, the China National Science Fund for Excellent Young Scholars in 2013, and the Changjiang Scholars Program from the Ministry of Education of China in 2017.

Shugen Wang received the B.S. degree in aerial photogrammetry from the Wuhan College of Surveying and Mapping, Wuhan, China, in 1984, the M.S. degree in photogrammetry and remote sensing from the Wuhan University of Surveying and Mapping Science and Technology, Wuhan, in 1994, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, in 2003.

He is currently a Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His major research interests include digital photogrammetry, high-spatial resolution remote sensing image processing, and computer vision.



Mingwei Sun was born in 1982. He received the B.S. degree from the Wuhan University of Technology, Wuhan, China, in 2004, and the Ph.D. degree from Wuhan University (WHU), Wuhan, in 2009.

He is currently an Associate Professor of photogrammetry and remote sensing with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include industrial and aerial photogrammetry, 3-D reconstruction of cultural relics and buildings, automatic ortho imagery mosaicking, true orthophoto production, and parallel computing.

Yangjun Ou was born in 1994. She received the B.S. degree in photogrammetry and remote sensing from Wuhan University (WHU), Wuhan, China, in 2016, where she is currently pursuing the Ph.D. degree in photogrammetry and remote sensing.

Her research interests include image retrieval, computer vision, and pattern recognition.

Yi Wan was born in 1991. He received the B.S. and the Ph.D. degrees from Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

He is currently pursuing his post-doctoral research with Wuhan University. His research interests include photogrammetry, computer vision, 3-D reconstruction, and change detection in remote sensing imagery.

Xiu Liu was born in 1981. He received the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2006 and 2012.

He is currently a Senior Engineer with the Beijing Institute of Space Mechanics and Electricity, Beijing. His research interests include remote sensing payload designed, image data processing, and optoelectronic imaging systems.