ISPRS Journal of Photogrammetry and Remote Sensing 163 (2020) 257–271

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

A differential information residual convolutional neural network for pansharpening

Menghui Jiang[a], Huanfeng Shen[a,c], Jie Li[b,*], Qiangqiang Yuan[b,c], Liangpei Zhang[c,d]

[a] School of Resource and Environmental Sciences, Wuhan University, Wuhan, China
[b] School of Geodesy and Geomatics, Wuhan University, Wuhan, China
[c] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan, China
[d] State Key Laboratory of Information Engineering, Survey Mapping and Remote Sensing, Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
Pansharpening
RCNN
Differential information mapping
Auxiliary gradient

ABSTRACT

Deep learning based methods are the state-of-the-art in panchromatic (PAN)/multispectral (MS) fusion (which is generally called "pansharpening"). In this paper, to solve the problem of the insufficient spatial enhancement in most of the existing deep learning based pansharpening methods, we propose a novel pansharpening method based on a residual convolutional neural network (RCNN). Differing from the existing deep learning based pansharpening methods that are mainly devoted to designing an effective network, we make novel changes to the input and the output of the network and propose a simple but effective mapping strategy. This strategy involves utilizing the network to map the differential information between the high spatial resolution panchromatic (HR-PAN) image and the low spatial resolution multispectral (LR-MS) image to the differential information between the HR-PAN image and the high spatial resolution multispectral (HR-MS) image, which is called the "differential information mapping strategy". Moreover, to further boost the spatial information in the fusion results, the proposed method makes full use of the LR-MS image and utilizes the gradient information of the up-sampled LR-MS image (Up-LR-MS) as auxiliary data to assist the network. Furthermore, an attention module and residual blocks are incorporated in the proposed network structure to maximize the ability of the network to extract features. Experiments on four data sets collected by different satellites confirm the superior performance of the proposed method compared to the state-of-the-art pansharpening methods.

1. Introduction

Due to the limitations of the sensor hardware, there are always trade-offs between the temporal resolution, spatial resolution, and spectral resolution in the captured remote sensing images (Shen et al., 2016). Remote sensing image fusion technologies that make full use of the complementary information between multi-source remote sensing images have been developed and widely used (Pohl and van Genderen, 1998; Sirguey et al., 2008; Martha et al., 2012). Pansharpening aims to integrate the complementary information of low-resolution multispectral (LR-MS) images and high-resolution panchromatic (HR-PAN) images. The pansharpening methods developed to date can be broadly classified into four major branches: 1) component substitution (CS)-based methods; 2) multiresolution analysis (MRA)-based methods; 3) variational model based methods; and 4) deep learning based methods (Meng et al., 2018).

The CS-based methods substitute the intensity component of the LR-MS image with that of the HR-PAN image. The classic CS-based methods include the intensity-hue-saturation (IHS) method (Carper et al., 2004), the Gram-Schmidt (GS) method (Laben and Brower, 2000), and the UNB PanSharp method (Zhang and Mishra, 2014), etc. These methods are the simplest and the most widely used, but they often suffer from spectral distortion. The MRA-based methods inject the high-frequency spatial details of the HR-PAN image into the LR-MS image. The representative MRA-based methods include decimated wavelet transform (DWT) method (Shahdoosti and Javaheri, 2017), undecimated wavelet transform method (Cheng et al., 2015), contourlets method (Nencini et al., 2007), and the generalized Laplacian pyramid with modulation transfer function (MTF) matched filter (MTF-GLP) method (Aiazzi et al., 2006). Compared with the CS-based methods, the MRA-based methods are generally sensitive to the spatial distortion, but produce less spectral distortion.

The variational model based methods regard the fusion process as an ill-posed inverse problem, and they construct the variational model

* Corresponding author.
E-mail address: jli89@sgg.whu.edu.cn (J. Li).

https://doi.org/10.1016/j.isprsjprs.2020.03.006
Received 18 September 2019; Received in revised form 6 February 2020; Accepted 4 March 2020
Available online 01 April 2020

based on Bayesian theory (Ballester et al., 2006; Zhang et al., 2012) or sparse expression theory (Jiang et al., 2014; Gogineni and Chaturvedi, 2018). Iterative optimization algorithms, such as the gradient descent algorithm (Moeller et al., 2008), the split Bregman iteration algorithm (Fang et al., 2013), or the alternating direction method of multipliers (ADMM) algorithm (Wei, 2015; Ghahremani et al., 2019), can be used to solve the constructed model to obtain the fusion result. The solid mathematical foundation of this kind of method can produce more precise fusion results than the former two categories. Unfortunately, the performance of this kind of method is largely dependent on the complex manually designed regularization terms, such as the total variation (TV) prior (Palsson et al., 2014), the nonlocal prior (Duran et al., 2017), and the Laplacian prior (Molina et al., 2008), which can greatly increase the computational complexity and make the model solving time-consuming.

Due to the powerful capability of feature extraction and learning (Zhang et al., 2017), deep learning has been shown to be effective at describing the nonlinear relationship between data. As such, deep learning based methods have been widely developed and applied in various fields (Shen et al., 2018; Wang et al., 2018; Xing et al., 2018; Zhang et al., 2018). Recently, many pansharpening methods based on deep learning have been proposed. For example, Masi et al. (2016) modified a super-resolution network and proposed a pansharpening neural network (PNN); Scarpa et al. (2018) proposed the target-adaptive CNN-based pansharpening method, which combines residual learning, L1 loss, and target-adaptive fine-tuning strategies to further improve the fusion accuracy; Wei et al. (2017) proposed the deep residual pansharpening neural network (DRPNN), which utilizes a deep convolutional filtering framework and residual network to boost the fusion accuracy; and Yuan et al. (2018) proposed the multiscale and multidepth convolutional neural network (MSDCNN), which utilizes filters with different sizes for extracting multiscale features.

Compared with the other categories, the deep learning based pansharpening methods can more easily achieve a better overall fusion accuracy. However, most of the existing deep learning based pansharpening methods have been adapted from single-image super-resolution (SISR) networks, and they ignore the biggest difference between pansharpening and SISR: the spatial details of SISR are inferred from the LR image, whereas the spatial details of pansharpening are extracted from the HR-PAN image. As a result, these methods are unable to make full use of the rich spatial information in the HR-PAN image, resulting in a blurred fusion result (Zhang et al., 2019). To alleviate this problem, some improved strategies have been proposed. For example, Liu et al. (2018) used a generative adversarial network (GAN) to boost the accuracy; He et al. (2019) incorporated the CS/MRA pansharpening categories into the network design, and proposed a detail injection based CNN; and Shen et al. (2019) combined deep learning and a variational model to make use of the complementarity of the two categories, to improve the spatial quality of the fusion results.

In this paper, a novel differential information residual convolutional neural network (DIRCNN) is proposed, which makes novel changes to the input and output of the network and effectively improves the spatial details of the fusion results. The main contributions of the proposed method are:

1) The differential information mapping strategy: Differing from the other pansharpening networks that are mainly devoted to designing an effective network, this strategy involves making novel changes to the input and output of the network. The network is utilized to learn the mapping from the differential information between the HR-PAN image and each band of the LR-MS image to the differential information between the HR-PAN image and each band of the ideal HR-MS image, which effectively enhances the spatial structure of the fusion results.
2) The auxiliary gradient strategy: The gradient of the LR-MS image contains rich spatial structure information of the desired HR-MS image. To further enhance the spatial details of the fusion results, the gradient information of the LR-MS image is added into the input of the network, to boost the fusion accuracy.
3) The combination of an attention module and residual blocks: To maximize the features extracted by the network, an attention module and residual blocks are combined in the network structure.

The rest of this paper is organized as follows. Section 2 introduces the novel strategies adopted in the proposed fusion method. Section 3 describes the detailed architecture of the proposed method. In Section 4, the experiments on both reduced-resolution and full-resolution images are presented, and both quantitative and visual assessments of the fusion results are provided and discussed. The conclusions and future prospects are summarized in Section 5.

2. Novel strategies of the proposed method

When designing a network for pansharpening, it is expected that the adopted network will effectively inject the spatial details of the HR-PAN image into each band of the LR-MS image, while maintaining the spectral relationship between bands of the LR-MS image. However, pansharpening networks usually have to make a trade-off between the two goals. Most of the existing CNN-based pansharpening methods (Masi et al., 2016; Wei et al., 2017; Scarpa et al., 2018; Yuan et al., 2018) usually perform well in maintaining the spectral relationship, but perform poorly in spatial enhancement (Zhang et al., 2019). To relieve this problem and boost the spatial details of the fusion results, three novel strategies are proposed in this paper: 1) the differential information mapping strategy; 2) the auxiliary gradient information strategy; and 3) the combination of an attention module and residual blocks. The first two strategies of the network input and output innovation are introduced below, and the third strategy of the network structure innovation is described in Section 3.

2.1. The differential information mapping strategy

The first strategy to be introduced is the novel differential information mapping strategy, which means mapping the differential information of the HR-PAN image and the LR-MS image to the differential information of the HR-PAN image and the HR-MS image through the network. The reason for selecting this strategy is described in detail below.

As displayed in Fig. 1(a), most of the existing CNN-based pansharpening methods (Masi et al., 2016; Wei et al., 2017; Scarpa et al., 2018; Yuan et al., 2018) usually utilize the network to learn the mapping from the observed LR-MS and HR-PAN images to the fused HR-MS image. The input of the network is {Up-LR-MS, HR-PAN}, where Up-LR-MS is the up-sampled LR-MS image that has the same spatial size as the HR-PAN image, and the output of the network is {HR-MS}. The fusion result, as displayed after the Test arrow in Fig. 1(a), shows that this kind of mapping strategy can achieve a satisfactory spectral fidelity; however, it suffers from insufficient spatial detail enhancement. The reason for this should be the underutilization of the HR-PAN image in the input in theory, because the spatial details of the desired high-resolution image only exist in the HR-PAN image (Zhang et al., 2019).

To make full use of the HR-PAN image and increase the spatial information of the fusion result, another mapping strategy is depicted in Fig. 1(b), which is to map the HR-PAN image to each band of the HR-MS image. Note that only multispectral images with four bands are discussed in this paper, so the input of the network in Fig. 1(b) is {HR-PAN, HR-PAN, HR-PAN, HR-PAN} and the output is still {HR-MS}. The fusion result, as displayed after the Test arrow in Fig. 1(b), shows that this kind of mapping strategy leads to rich spatial enhancement, but very serious spectral distortion, which may be due to the absence of the LR-MS image in the input.

To make use of the complementarity of the two strategies in

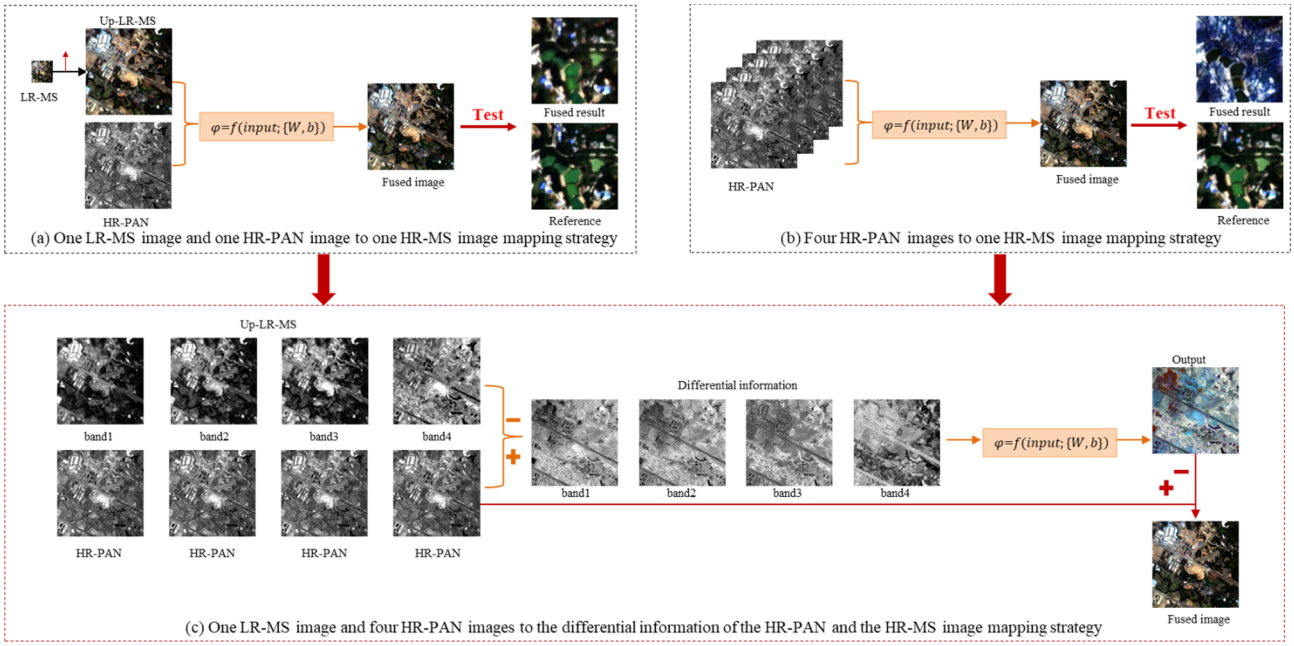

Fig. 1. The different CNN-based pansharpening frameworks.

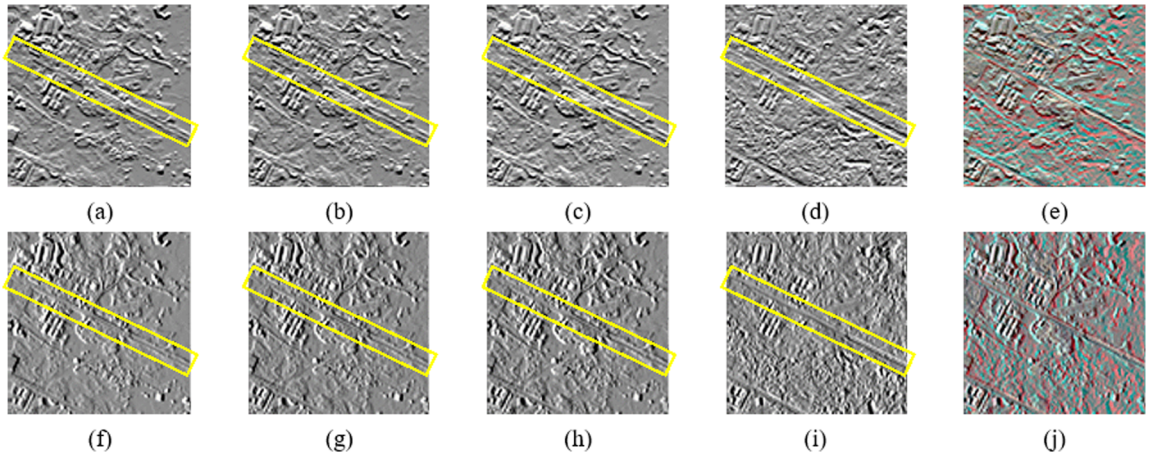

Fig. 2. Horizontal and vertical gradient of the Up-LR-MS image. (a)–(d) Horizontal gradient of the first to the fourth band of the Up-LR-MS image. (e) False-color synthesis of the horizontal gradient image (i.e. 4-3-2 band combination). (f)–(i) Vertical gradient of the first to the fourth band of the Up-LR-MS image. (j) False-color synthesis of the vertical gradient image (i.e. 4-3-2 band combination).

Fig. 1(a–b), it is natural to assign the HR-PAN image to each band of the LR-MS image. The simplest way is to concatenate the images in the spectral dimension, as shown in the left of Fig. 1(c); however, such input settings are redundant, and the increase of the input image number increases the computational complexity of the network. To optimize it, the differential information mapping strategy is proposed, which is to use the differential information between the HR-PAN image and each band of the LR-MS image as the input of the network. As depicted in Fig. 1(c), the input of the network is {HR-PAN - LR-MS$_i$, …}$_{i=1:4}$ and the output is {HR-PAN - HR-MS$_i$, …}$_{i=1:4}$, where LR-MS$_i$ denotes the ith band of the LR-MS image. There are three advantages to the differential information mapping strategy:

Firstly, when we assign the HR-PAN image to each band of the LR-MS image, the multiple use of the HR-PAN image provides sufficient spatial information to meet the spatial enhancement needs of each band of the desired HR-MS image. Secondly, compared with the simple concatenation strategy, the decrease of the input image number in the proposed differential information mapping strategy reduces the computational complexity of the network. Lastly, but most importantly, the input of the network in Fig. 1(c) contains more information than the output, which differs from most of the pansharpening networks in the literature. As introduced above, the input of the network in Fig. 1(c) is {HR-PAN - LR-MS$_i$, …}$_{i=1:4}$ and the output is {HR-PAN - HR-MS$_i$, …}$_{i=1:4}$. Since the HR-MS image contains richer information than the LR-MS image, the information in the output is less than that in the input. As is well known, when the network maps less information to more information, the function is to create information, but when the network maps more information to less information, the function is to refine the information. Intuitively, it seems much more feasible for the network to perform the task of refining information than the task of creating information.

2.2. The auxiliary gradient information strategy

In remote sensing images, the spatial structure refers to the spatial composition and arrangement of pixels, which is reflected by the spatial changes of pixel values. The gradient information of the image, i.e. the difference between adjacent pixels, can directly represent the spatial

structure of the image. Therefore, to further increase the spatial details of the fusion results, the second strategy to be introduced is the auxiliary gradient information strategy, which is to utilize the gradient information of the Up-LR-MS image to assist the network. Fig. 2 shows the gradient of the Up-LR-MS image, where the first line represents the horizontal gradient and the second line represents the vertical gradient. In Fig. 2(a–d) and (f–i), it can be seen that the gradient image of each band contains rich edge structure information of the desired HR-MS image, as shown in the road edge inside the yellow rectangle. Moreover, the color information shown in the false-color synthesis image in Fig. 2(e) and (j) indicates that the gradient image also contains some spectral information of the desired HR-MS image. Thus, using the gradient information of the Up-LR-MS image as auxiliary data not only improves the spatial details of the fusion results, but also alleviates the possible spectral distortion.

3. The proposed method

3.1. Details of the proposed pansharpening framework

Before describing the proposed pansharpening framework in detail, it is necessary to introduce the notations used in this paper. $\boldsymbol{X} \in \mathbb{R}^{M\times N\times S}$ denotes the ideal HR-MS image, where M, N, and S represent the width, height, and band number of the image, respectively. $\boldsymbol{Y} \in \mathbb{R}^{m\times n\times S}$ denotes the observed LR-MS image, and $\boldsymbol{Z} \in \mathbb{R}^{M\times N\times 1}$ denotes the observed HR-PAN image. M /m = N /n is the spatial resolution ratio of the LR-MS image to the HR-MS image.

Fig. 3 is the flowchart of the proposed method. As shown in Fig. 3, the LR-MS image is first bicubic-up-sampled to the same spatial size as the corresponding HR-PAN image to obtain the Up-LR-MS image. The HR-PAN image is then histogram matched by the Up-LR-MS image to obtain the histogram-matched HR-PAN image (HHR-PAN) as follows:

$$\widehat{\boldsymbol{z}}_i = \frac{\boldsymbol{Z} - \mu(\boldsymbol{Z})}{std(\boldsymbol{Z})} \cdot std(\widetilde{\boldsymbol{y}}_i) + \mu(\widetilde{\boldsymbol{y}}_i) \tag{1}$$

where $\widetilde{\boldsymbol{Y}} = [\widetilde{\boldsymbol{y}}_1, \cdots \widetilde{\boldsymbol{y}}_i \cdots]_{i=1,\cdots,S}$ is the Up-LR-MS image, and $\widetilde{\boldsymbol{y}}_i$ represents the ith band of the Up-LR-MS image. $\mu(\widetilde{\boldsymbol{y}}_i)$ and $std(\widetilde{\boldsymbol{y}}_i)$ denote the mean and the standard deviation of $\widetilde{\boldsymbol{y}}_i$, respectively. $\widehat{\boldsymbol{Z}} = [\widehat{\boldsymbol{z}}_1, \cdots \widehat{\boldsymbol{z}}_i \cdots]_{i=1,\cdots,S}$ is the HHR-PAN image calculated through Eq. (1). The histogram match (Aiazzi et al., 2006; Rahmani et al., 2010; Vivone et al., 2014) is used to eliminate the spectral distortion caused by the gray value difference between the Up-LR-MS image and the HR-PAN image.

As displayed in Fig. 3, the proposed network has two inputs, which are made up of two types of data. The first type is the differential information between the HHR-PAN image and the Up-LR-MS image, as shown in the red border in Fig. 3, and the second type is the gradient information of the Up-LR-MS image, as shown in the green border in Fig. 3. The two inputs of the proposed network can be denoted as follows:

$$\boldsymbol{D}^{in1} = \widehat{\boldsymbol{Z}} - \widetilde{\boldsymbol{Y}} \tag{2}$$

$$\boldsymbol{D}^{in2} = \{\boldsymbol{D}_i^{in1}, \boldsymbol{g}_i, \cdots\}_{i=1,\cdots,S} \tag{3}$$

where $\boldsymbol{D}^{in1} \in \mathbb{R}^{M\times N\times S}$ represents the first set of input data, which is the differential information of the HHR-PAN image and the Up-LR-MS image. $\boldsymbol{D}^{in2} \in \mathbb{R}^{M\times N\times 3S}$ represents the second set of input data, which is the concatenation of the differential information and the gradient information in the spectral dimension. $\boldsymbol{g}_i = [\boldsymbol{g}_i^h, \boldsymbol{g}_i^v] \in \mathbb{R}^{M\times N\times 2}$ in Eq. (3) denotes the horizontal and vertical gradient information of the ith band of the Up-LR-MS image.

The desired output of the proposed network is assumed to be the differential information of the HHR-PAN image and the ideal HR-MS image, which can be written as:

$$\boldsymbol{D}^{lab} = \widehat{\boldsymbol{Z}} - \boldsymbol{X} \tag{4}$$

where $\boldsymbol{D}^{lab} \in \mathbb{R}^{M\times N\times S}$ represents the label data that supervises the network training.

The final fusion result of the proposed method can be obtained as follows:

$$\boldsymbol{D}^{fused} = \widehat{\boldsymbol{Z}} - f((\boldsymbol{D}^{in1}, \boldsymbol{D}^{in2}); \Theta) \tag{5}$$

where $\boldsymbol{D}^{fused} \in \mathbb{R}^{M\times N\times S}$ is the fused HR-MS image, $f(\cdot)$ denotes the network, Θ denotes the trainable parameters of the network, and $f((\boldsymbol{D}^{in1}, \boldsymbol{D}^{in2}); \Theta)$ is the output of the network.

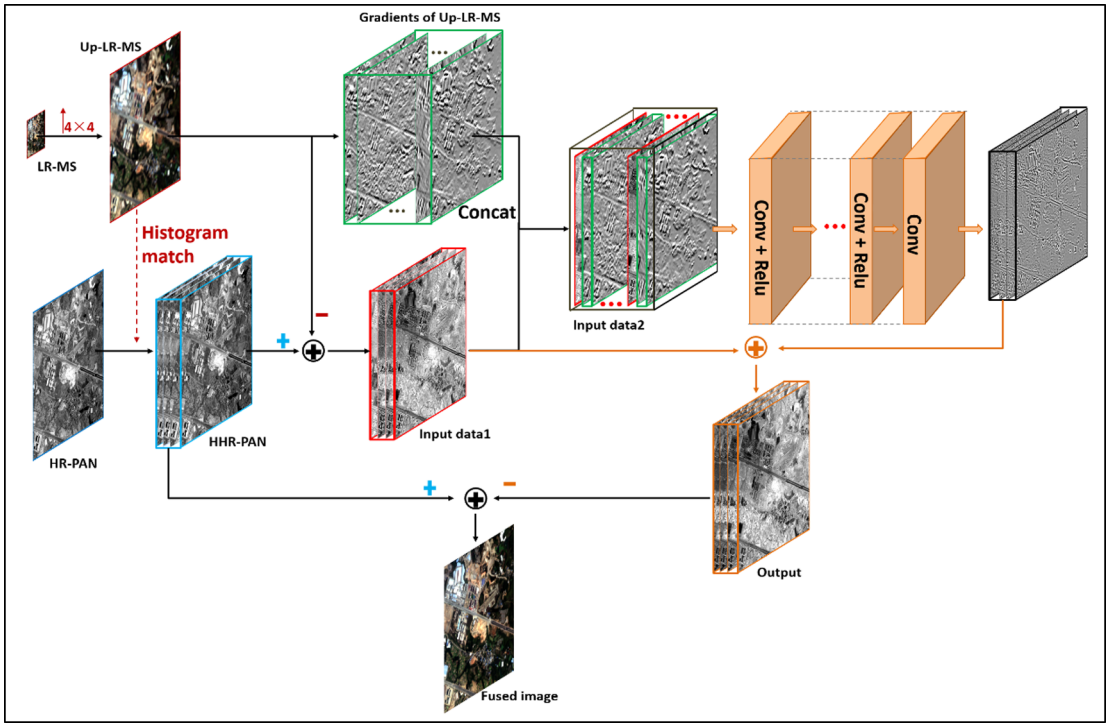

Fig. 3. Flowchart of the proposed method.

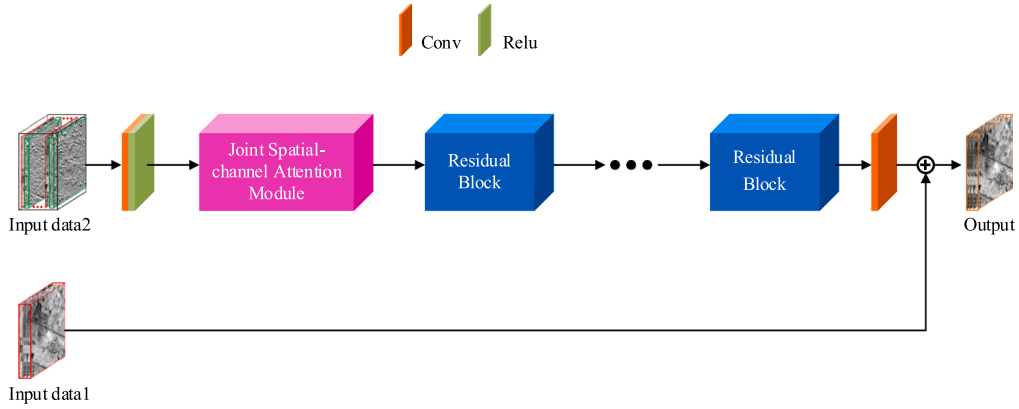

Fig. 4. Overview structure of the proposed network.

3.2. Architecture of the proposed network

Fig. 4 gives a brief overview of the structure of the proposed network. As shown in Fig. 4, the proposed network has four types of blocks. The first is a "Conv + ReLU" block, the second is a "joint spatial-channel attention module" block, followed by four "residual block" units, and the last is a "Conv" block. Specifically, the first block includes a convolutional layer and a rectified linear unit (ReLU) activation function layer, where the convolutional layer consists of 64 filters of $3 \times 3 \times 3S$. S denotes the band number of the MS image, which is fixed at 4 in this paper. The last block includes only one convolutional layer, which consists of S filters of $3 \times 3 \times 64$. To maximize the ability of the network to extract features, we combine the popular attention module known as the "joint spatial-channel attention module" and the popular residual learning units known as "residual block" in the middle of the network structure. Details of the attention module and residual blocks are provided below.

3.2.1. The joint spatial-channel attention module

The attention mechanism, which has the ability to recalibrate the extracted feature maps, is a popular network structure applied in various computer vision problems, such as classification (Hu et al., 2018), semantic segmentation (Fu et al., 2018), and super-resolution (Kim et al., 2018). Fig. 5 shows the detailed structure of the adopted spatial-channel attention module, which is made up of a channel attention mechanism and a spatial attention mechanism. As displayed in Fig. 5, the channel attention mechanism consists of an average pooling (Avgpool) layer and two fully connected (FC) layers. In the channel attention mechanism branch, the Avgpool layer is first adopted to extract the global statistical features of each input feature map, and then the two FC layers are utilized to generate the channel features. The spatial attention mechanism consists of three grouped convolutional (Gconv) layers (Krizhevsky et al., 2012; Xie et al., 2017). In each Gconv layer, the input feature maps are first divided into 64 groups, i.e., each input feature map forms a group independently, and then 64 filters of size $3 \times 3 \times 1$ are utilized to extract the spatial feature map of each group.

The spatial feature maps extracted from the spatial attention mechanism and the channel feature maps generated from the channel attention mechanism are then combined through a scale layer and a sigmoid layer. The output of the spatial-channel attention module can be written as follows:

$$\mathbf{F}_{output} = \mathbf{F}_{input} \oplus (\mathbf{F}_{input} \otimes \sigma(Scale(\mathbf{M}_C, \mathbf{M}_S))) \quad (6)$$

where $\mathbf{F}_{input}$ and $\mathbf{F}_{output}$ denote the input feature maps and the output feature maps of the spatial-channel attention module, respectively. $\mathbf{M}_C$and $\mathbf{M}_S$ represent the generated channel features and spatial features, respectively. $Scale(\cdot)$ and $\sigma(\cdot)$ represent the scale function and the sigmoid function, respectively. $\oplus$ and $\otimes$ are element-wise addition and element-wise multiplication functions, respectively.

3.2.2. Residual block

To train a more accurate network, we adopt the popular residual learning strategy proposed by He et al. (2016) to deal with the problem of training accuracy degradation with increasing network depth. The effectiveness of the residual learning strategy has been verified in many tasks (Kiku et al., 2013; Timofte et al., 2014). In the proposed framework, the residual learning is utilized in two ways. The first is to add the input data1 to the feature maps generated by the last convolutional

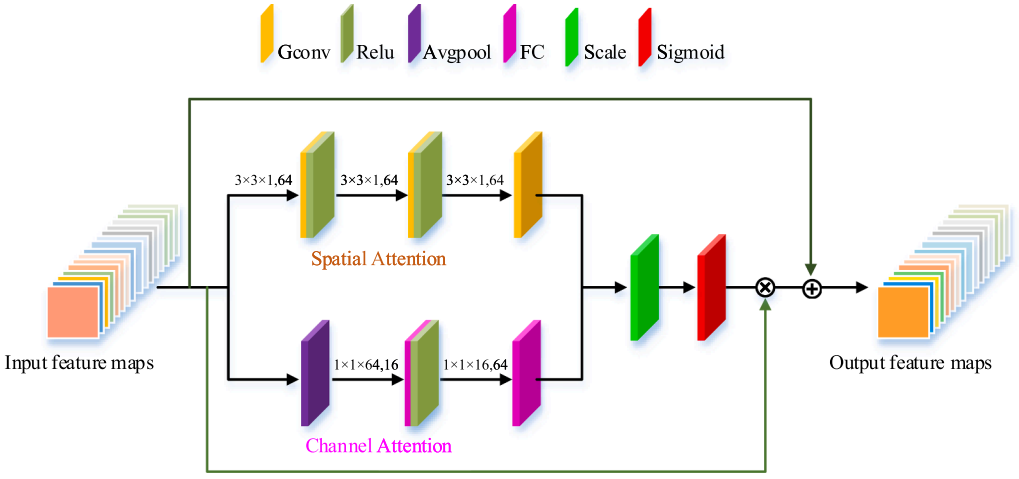

Fig. 5. The joint spatial-channel attention module.

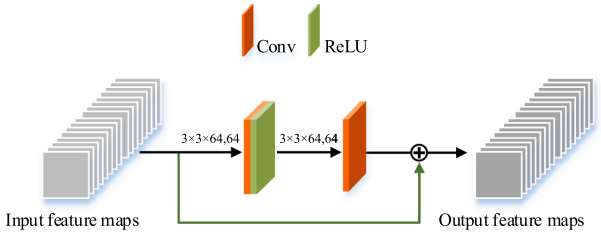

Fig. 6. Residual block.

layer, as shown in Fig. 4, which is to transfer the information of the input data1 to the output directly. The second is to use the residual learning in the residual block structure, as shown in Fig. 6.

As shown in Fig. 6, the residual block has two convolutional layers, of which the first convolutional layer is followed by the ReLU function layer, and 64 filters of size 3 × 3 × 64 are utilized in each convolutional layer. ⊕ is a pixel-by-pixel addition function, which is also known as a "skip connection". The input of a residual block is added to the feature maps obtained by the second convolutional layer to generate the output of the residual block. The residual learning utilized in the residual block cascades the low-level features and the high-level features, which reduces the loss of the useful information in the feature extraction process.

4. Experiments and discussion

4.1. Comparison methods and quantitative evaluation indices

To verify the effectiveness of the proposed method, experiments on both reduced-resolution images (also called simulated experiments) and full-resolution images (also called real-data experiments) are conducted on various data sets. Six mainstream fusion algorithms belonging to different branches are selected for the comparison: adaptive intensity-hue-saturation (AIHS) (Rahmani et al., 2010); matting-model pansharpening (MMP) (Kang et al., 2013) of the CS-based class; MTF-GLP (Aiazzi et al., 2006) of the MRA-based category; coupled nonnegative matrix factorization (CNMF) (Yokoya et al., 2011) and two-step sparse coding (TSSC) (Jiang et al., 2014) of the model-based branch; and the deep residual pansharpening neural network (DRPNN) (Wei et al., 2017) of the CNN-based class. Note that the network of the DRPNN method is retrained with the data sets used in this study. All the experiments are implemented in MATLAB (R2016a).

The fusion results of the simulated experiments are evaluated using the popular Wald's protocol (Wald et al., 1997). According to the protocol, the observed LR-MS and HR-PAN images are first down-sampled by their spatial resolution ratio with bicubic interpolation. The fusion methods are then performed on the down-sampled MS and PAN images, with the original LR-MS image used as the ground-truth reference. A number of different indices have been proposed to precisely evaluate the spectral and spatial performance of the fusion results, and six representative indices are utilized in this study. They are: the relative dimensionless global error in synthesis (ERGAS) (Vivone et al., 2014), the spectral angle mapper (SAM) (Vivone et al., 2014), the Q metric (Vivone et al., 2014), the peak-signal-to-noise ratio (PSNR), the structural similarity index (SSIM) (Wang et al., 2004), and the spatial correlation coefficient (SCC) (Zhou et al., 1998). Among them, SAM is a spectral quality metric, SSIM and SCC are spatial quality metrics, and ERGAS, Q, and PSNR are comprehensive spatial-spectral quality metrics.

In the real-data experiments, the pansharpening methods are directly performed on the observed LR-MS and HR-PAN images, so that there is no reference image. In this case, the quality with no reference (QNR) index (Vivone et al., 2014) (which is composed of a spectral distortion (D_λ) index and a spatial distortion (D_S) index), the entropy, and the spatial frequency (SF) are utilized to evaluate the fusion products.

4.2. Data sets for training and testing

Data sets from four satellites are used in this study: QuickBird, IKONOS, Gaofen-2 (GF-2), and Gaofen-1 (GF-1). Due to the limited amount of data available, there is not enough data to train the network for each satellite. Thus, only images of the QuickBird satellite and GF-2 satellite are utilized in the network training, and images of all four satellites are utilized in the testing. The gray values of all the images are normalized to [0,1].

1) **Training data sets:** The network models are trained on the QuickBird images and the GF-2 images.
   a) When training the QuickBird network, three LR-MS and HR-PAN image pairs are utilized to generate the training patches. As detailed in Table 1, the first pair of images is of Nanchang, Jiangxi province, China, with the size of 5200 × 4400 × 4 for the LR-MS image and 20,800 × 17,600 × 1 for the HR-PAN image, for which the resolution are 2.44 m for the LR-MS image and 0.61 m for the HR-PAN image, and the center location is 115.92E, 28.65N. Most of the landscape in this study area is buildings, followed by vegetation. The second pair of images is of Shenzhen, Guangdong province, China, and the third pair of images is of Yichang, Hubei province, China. The main types of landscape in these two study areas are buildings, vegetation, and water, of which buildings and vegetation are the most common. In total, 58,432 patches of size 31 × 31 are randomly generated from these three image pairs for the QuickBird network.
   b) When training the GF-2 network, two LR-MS and HR-PAN image pairs are utilized to generate the training patches. The two pairs of images are both of Nanning, Guangxi province, China, with different locations, as detailed in Table 1. In total, 58,752 patches of size 31 × 31 are randomly generated from the two Nanning image pairs for the GF-2 network.
2) **Test data sets:** Four data sets of different satellites are employed in

Table 1
Attributes of the training data sets.

Training data sets	Region	Size	Resolution	Location	Landscape	Bach size	Patch size	Patch number
QuickBird	Nanchang	5200 × 4400 × 4(MS), 20,800 × 17,600 × 1(PAN)	2.44 m(MS), 0.61 m(PAN)	115.92E, 28.65N	buildings, vegetation	64	31 × 31	58,432
	Shenzhen	1000 × 1000 × 4, 4000 × 4000 × 1	2.44 m, 0.61 m	114.06E, 22.56N	buildings, vegetation	64	31 × 31	
	Yichang	3600 × 1000 × 4, 14,400 × 4000 × 1	2.44 m, 0.61 m	111.30E, 30.68N	vegetation, buildings	64	31 × 31	
GF-2	Nanning	7246 × 7059 × 4, 29,289 × 28,516 × 1	4 m, 1 m	108.32E, 22.80N	buildings, vegetation	64	31 × 31	58,752
	Nanning	7448 × 7083 × 4, 31,085 × 29,573 × 1	4 m, 1 m	108.48E, 22.86N	vegetation, buildings,	64	31 × 31	

Table 2
Attributes of the test data sets.

Test data sets	Satellites	Region	Size	Sub_size	Number	Resolution	Location	Landscape
Simulated experiments	QuickBird	Wuhan	1000 × 1000 × 4 4000 × 4000 × 1	250 × 250 × 4 1000 × 1000 × 1	16	2.44 m, 0.61 m	114.36E, 30.53N	buildings, vegetation
	GF-2	Nanning	2000 × 1600 × 4 8000 × 6400 × 1	400 × 400 × 4 1600 × 1600 × 1	20	4 m, 1 m	108.50E, 22.93N	vegetation, buildings
Real-data experiments	IKONOS	Wuhan	400 × 400 × 4 1600 × 1600 × 1	200 × 200 × 4 800 × 800 × 1	4	4 m, 1 m	114.19E, 30.51N	vegetation, buildings
	GF-1	Huizhou	400 × 400 × 4 1600 × 1600 × 1	200 × 200 × 4 800 × 800 × 1	4	9.6 m, 2.4 m	114.35E, 23.04N	buildings, vegetation

the experiments, as follows. Note that all the test data sets are spatially disjoint from the patches used in the network training.

a) Table 2 details the test data sets. The first data set used in the simulated QuickBird experiments is a pair of images of Wuhan, Hubei province, China, with the size of 1000 × 1000 × 4 for the LR-MS image and 4000 × 4000 × 1 for the HR-PAN image. It is worth mentioning that, considering the fusion efficiency of all the comparison methods (especially the model-based methods), the images are divided into 16 pairs of small images with the size of 250 × 250 × 4 for each LR-MS image and 1000 × 1000 × 1 for the HR-PAN image.
b) The second data set used in the simulated GF-2 experiment is a pair of images of Nanning, Guangxi province, China, with the size of 2000 × 1600 × 4 and 8000 × 6400 × 1, respectively. The images are divided into 20 pairs of small images with the size of 400 × 400 × 4 and 1600 × 1600 × 1, respectively.
c) The third data set used in the real-data IKONOS experiments is a pair of images of Wuhan, Hubei province, China, with the size of 400 × 400 × 4 and 1600 × 1600 × 1, respectively. These images are divided into four pairs of images with the size of 200 × 200 × 4 and 800 × 800 × 1, respectively.
d) The last data set is used in the real-data GF-1 experiments is a pair of images of Huizhou, Guangxi province, China, with the size of 400 × 400 × 4 and 1600 × 1600 × 1. These images are divided into four pairs of test images with the size of 200 × 200 × 4 and 800 × 800 × 1, respectively.

4.3. Comparison of the accuracy of DIRCNNs with different strategies

To verify the effectiveness of the main strategies of the proposed method, three networks are trained on the same data with different strategies, as listed in Table 3. The first network is a simple differential information residual CNN that uses neither the attention module and residual blocks nor the auxiliary gradient information, which is denoted as “DIRCNN-plain”. Note that the non-use of residual blocks means the removal of the skip connection function in the residual block structure displayed in Fig. 6. The second network uses the attention module and residual blocks, but not the auxiliary gradient information, which is denoted as “DIRCNN-nog”. The third network uses the attention module and residual blocks, as well as the auxiliary gradient information, and is denoted as DIRCNN.

Fig. 7 compares the training loss curves of the different networks trained on GF-2 images. Firstly, by comparing the green line and the blue line, it can be seen that, at the beginning of the iterations, the loss value of DIRCNN-nog is higher than that of DIRCNN-plain, but with the increase of the iteration, the final convergence value of DIRCNN-nog is slightly lower than that of DIRCNN-plain. This indicates the effectiveness of utilizing the attention module and the residual blocks in the network structure. Next, by comparing the blue line and the red line, it can be observed that the iterative convergence value of DIRCNN is much lower than that of DIRCNN-nog, which demonstrates the effectiveness of the proposed auxiliary gradient strategy.

Table 3
Networks with different strategies.

	Attention module and residual block	Auxiliary gradient data
DIRCNN-plain	×	×
DIRCNN-nog	√	×
DIRCNN	√	√

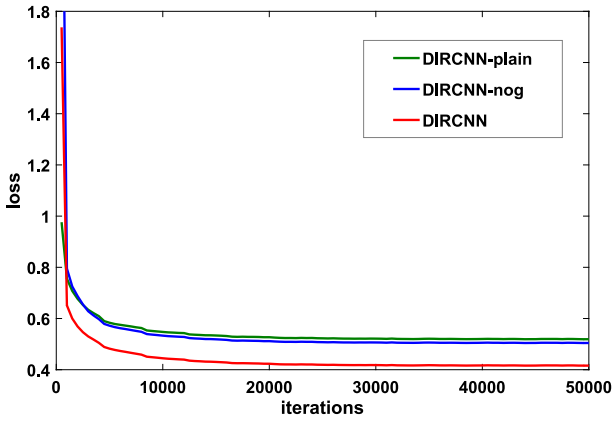

Fig. 7. Training loss curves of the different networks.

To further analyze the accuracy of the three networks, the test data for the simulated GF-2 experiments described in Section 4.2 are selected for both qualitative and quantitative evaluation. Table 4 lists the quantitative evaluation results, with the average for the 20 groups. In Table 4, the best performance for each index is marked in red, and the second-best performance for each quality index is marked in blue. As can be seen in Table 4, DIRCNN performs the best in all the indices, and DIRCNN-nog performs slightly better than DIRCNN-plain in all the indices, which is consistent with the convergence of the training loss curves of the three networks in Fig. 7.

Table 4
Quantitative results of the simulated GF-2 images (16 groups)

Algorithm	Ideal data	DIRCNN-plain	DIRCNN-nog	DIRCNN
ERGAS	**0**	1.9586	1.9328	1.8098
SAM	**0**	2.5978	2.5471	2.4412
Q	**1**	0.9476	0.9497	0.9557
PSNR	$+\infty$	36.3103	36.4519	36.9075
$SSIM_B$	**1**	0.9906	0.9912	0.9926
$SSIM_G$	**1**	0.9913	0.9915	0.9936
$SSIM_R$	**1**	0.9885	0.9888	0.9913
$SSIM_{NIR}$	**1**	0.8844	0.8886	0.9004
$SSIM_{AVG}$	**1**	0.9637	0.9650	0.9695
SCC_B	**1**	0.9215	0.9264	0.9323
SCC_G	**1**	0.9188	0.9218	0.9345
SCC_R	**1**	0.8987	0.8991	0.9131
SCC_{NIR}	**1**	0.6218	0.6361	0.6626
SCC_{AVG}	**1**	0.8402	0.8458	0.8606

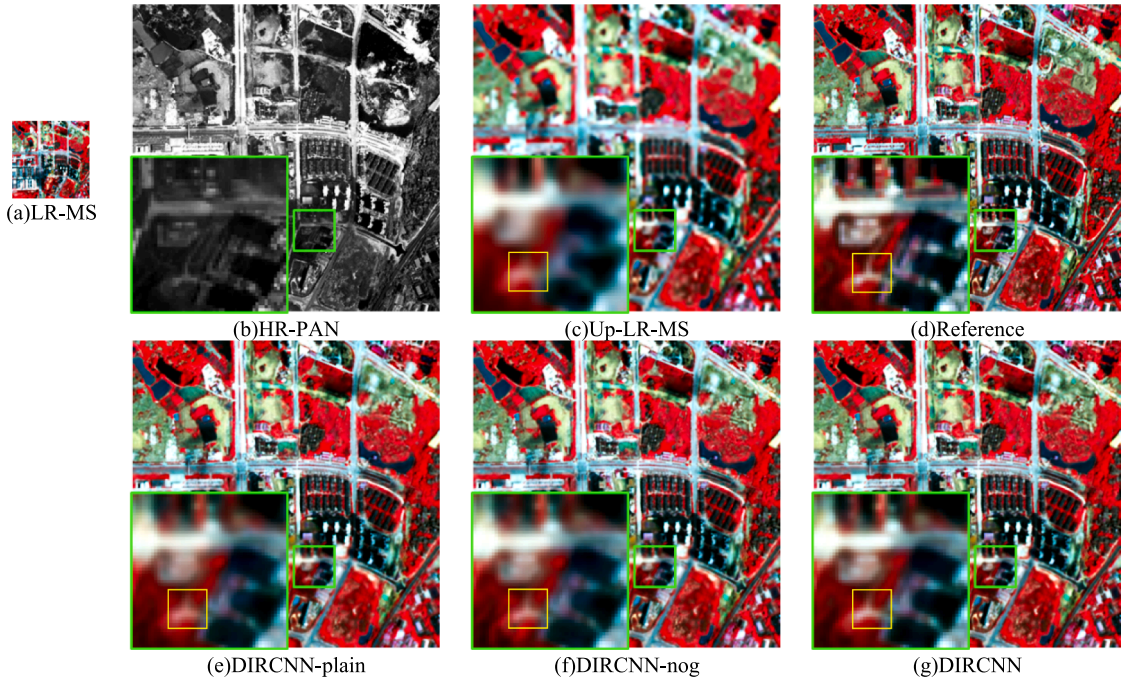

Fig. 8. Simulated QuickBird fusion results of three DIRCNNs.

Fig. 8 displays a group of fusion results of the three networks in false-color synthesis (i.e. 4-3-2 band combination), where the lower-left corner is a magnified display of the image inside the green rectangle. Fig. 8(a–b) show the LR-MS and HR-PAN images to be fused. Fig. 8(c) is the Up-LR-MS image, which is generated by bicubic interpolation of the LR-MS image in Fig. 8(a). Fig. 8(d) is the reference image, and Fig. 8(e–f) are the fusion results. Spectrally, the color information of the three fusion results is consistent with the reference image, and no significant color distortion is found. Spatially, compared with the Up-LR-MS image, the spatial information of the three fusion results is clearer, as can be seen in the vegetation area in the upper-right corner of the images in Fig. 8(c–g). Moreover, in the zoomed area in the lower-left corner, among the three fusion results, it can be found that the result of DIRCNN is richer in spatial details than the results of DIRCNN-plain and DIRCNN-nog. Little difference can be found between the fusion results of DIRCNN-plain and DIRCNN-nog, but if the white junction in the yellow rectangle area is examined closely, it can be seen that DIRCNN-nog is slightly clearer than DIRCNN-plain, DIRCNN is clearer than both DIRCNN-nog and DIRCNN-plain.

In summary, the effectiveness of the three DIRCNN networks show the feasibility of the proposed differential information mapping strategy. Moreover, from the loss curves, qualitative and quantitative assessments, it is found that DIRCNN-nog performs slightly better than DIRCNN-plain, which demonstrates the effectiveness of the proposed attention module and residual blocks. DIRCNN performs much better than the other two networks, which confirms the effectiveness of the proposed auxiliary gradient information strategy. In the following, we compare the effects of the proposed DIRCNN and the other state-of-the-art methods.

4.4. Simulated experiments

4.4.1. Simulated QuickBird experiments

The first series of simulated experiments are performed on the QuickBird images. Table 5 lists the quantitative evaluation results of the simulated QuickBird experiments, with the average of 16 groups, where columns 3–7 list five non-CNN-based methods and columns 8–9 list two CNN-based approaches. In Table 5, the best performance for each index is marked in red, the second-best performance for each quality index is marked in blue, and the third-best performance for each quality index is marked with underline. Among all the comparison methods, the two CNN-based methods outperform the other methods in all the quality indices, which shows the outstanding learning ability of CNNs. Moreover, by comparing the results of DRPNN and DIRCNN, it can be observed that DIRCNN performs better than DRPNN in all the indices, which indicates that the proposed DIRCNN is superior to DRPNN in both spectral fidelity and spatial enhancement in this series of experiments.

A group of simulated QuickBird fusion images is selected to be displayed in Fig. 9 in false-color synthesis (i.e. 4-3-2 band combination), where the lower-right corner is a magnified display of the image inside the green rectangle. From Fig. 9, the visual images of AIHS, MMP, MTF-GLP, CNMF, and TSSC show obvious spectral distortion. In more detail, the fusion results of MMP and MTF-GLP show very poor spectral preservation, as can be seen in the zoomed area in Fig. 9(b–c). AIHS not only produces severe spectral distortion, but also obvious blurring, as shown in the edges of the buildings in the zoomed area in Fig. 9(a). TSSC exhibits some spatial aliasing in the fusion result, as can be seen in the area between the three buildings in the zoomed area of Fig. 9(e). Among the five non-CNN-based methods, the result of CNMF is the closest to the reference image, visually. However, CNMF exhibits poor color contrast, for example, the color of the buildings in the zoomed area of Fig. 9(d) is orange, while that in the zoomed area of Fig. 9(h) is white. Compared with the five non-CNN-based methods, the color of the two CNN-based approaches are more consistent with that of the reference image, as can be seen in the vegetation inside the yellow rectangle in Fig. 9(f–h). Moreover, comparing the vegetation inside the yellow rectangle, it can be found the vegetation of DIRCNN is closer to that of the reference image; and comparing the edges of the three buildings below the yellow rectangle, it can be found that the building edges of DRPNN show a slight halo effect, whereas the building edges of DIRCNN and the reference image are visually clearer.

To further analyze the effects of the different methods, Fig. 10 displays boxplots of the absolute difference images between the various fusion results and the reference of Fig. 9 in each band. The smallest mean and median, the largest minimum, the smallest maximum, and the smallest box region from the 25th to the 75th or the 1st to the 99th percentile in Fig. 10(a–d) show that the fusion result of the proposed method is the closest to the reference image.

Table 5
Quantitative results of the simulated quickbird images (16 groups)

Algorithm	Ideal data	AIHS	MMP	MTF-GLP	CNMF	TSSC	DRPNN	DIRCNN
ERGAS	**0**	3.6916	3.2882	3.5496	3.2661	3.2180	2.5885	2.3580
SAM	**0**	3.7683	3.5055	3.6767	3.2002	3.3688	2.6385	2.5019
Q	**1**	0.8715	0.8944	0.8863	0.9060	0.9109	0.9319	0.9432
PSNR	$+\infty$	33.2135	34.4092	33.9369	34.3349	34.5604	36.2936	37.1676
$SSIM_B$	**1**	0.9418	0.9369	0.9251	0.9455	0.9403	0.9585	0.9651
$SSIM_G$	**1**	0.8869	0.8911	0.8825	0.9087	0.8947	0.9230	0.9400
$SSIM_R$	**1**	0.8746	0.8781	0.8675	0.9058	0.8855	0.9270	0.9401
$SSIM_{NIR}$	**1**	0.8029	0.8842	0.8838	0.8878	0.8858	0.9100	0.9313
$SSIM_{AVG}$	**1**	0.8766	0.8976	0.8897	0.9120	0.9016	0.9296	0.9441
SCC_B	**1**	0.7855	0.7474	0.5682	0.7729	0.7284	0.8566	0.8671
SCC_G	**1**	0.7673	0.6997	0.5903	0.7505	0.6973	0.8403	0.8632
SCC_R	**1**	0.4762	0.4489	0.4009	0.5223	0.4602	0.6552	0.6893
SCC_{NIR}	**1**	0.6252	0.6105	0.6129	0.6365	0.6024	0.7573	0.8383
SCC_{AVG}	**1**	0.6635	0.6266	0.5431	0.6705	0.6221	0.7774	0.8145

4.4.2. Simulated GF-2 experiments

The second series of simulated experiments are performed on the GF-2 images. Table 6 lists the quantitative evaluation results of the simulated GF-2 experiments, with the average of 20 groups, where the best result is marked in red, the second-best result is marked in blue, and the third-best result is marked with underline. As displayed in Table 6, the two CNN-based methods outperform the five non-CNN-based methods in all the indices, and for the five non-CNN-based methods, MMP and TSSC perform better. Moreover, for the two CNN-based methods, the proposed DIRCNN performs much better than DRPNN in all the indices.

Fig. 11 displays a group of simulated GF-2 fusion results in false-color synthesis (i.e. 4-3-2 band combination). By comparing the fusion results, it can be observed that AIHS, MMP, and MTF-GLP all perform poorly with regard to spatial detail enhancement. Specifically, compared with the Up-LR-MS image, the spatial information in the results of AIHS, MMP, and MTF-GLP is effectively increased in the bare soil part of the zoomed area, but in the vegetation area of the zoomed area, no significant increase in spatial information can be seen. Comparing the zoomed area of CNMF in Fig. 11(d) with that of the reference image

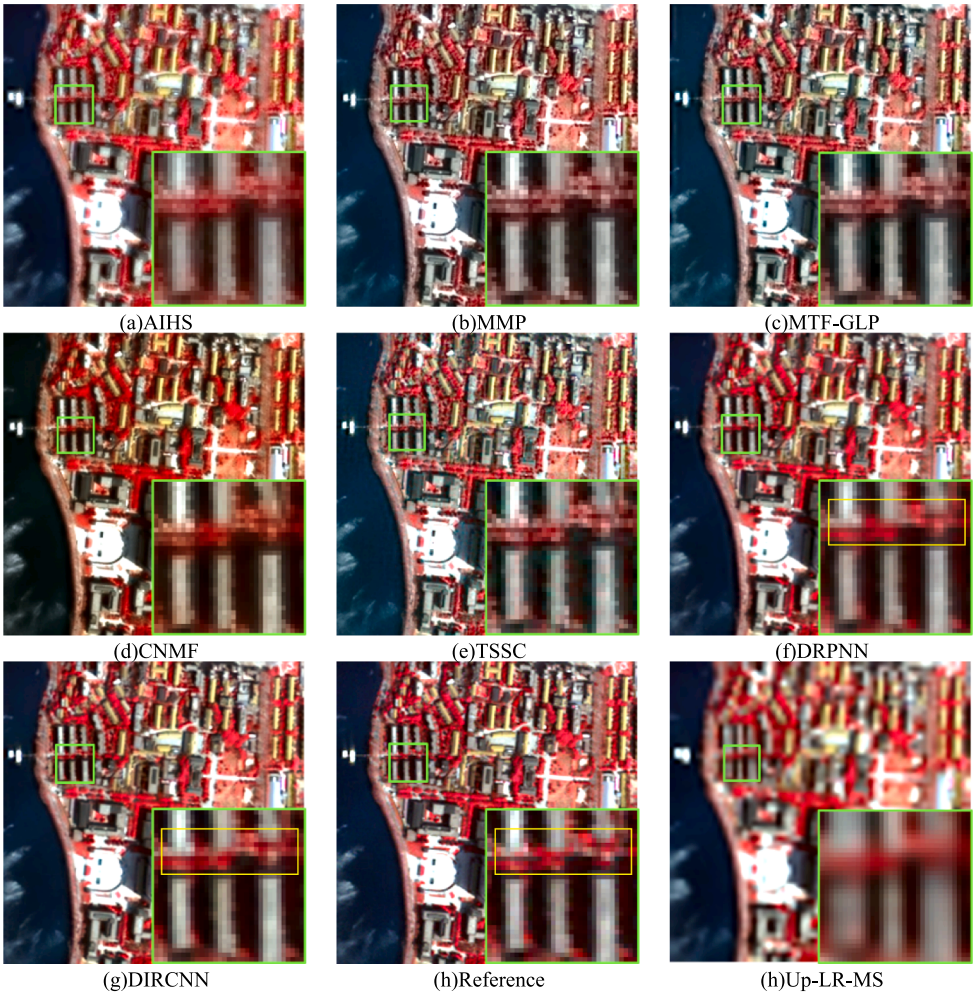

Fig. 9. Simulated QuickBird fusion results.

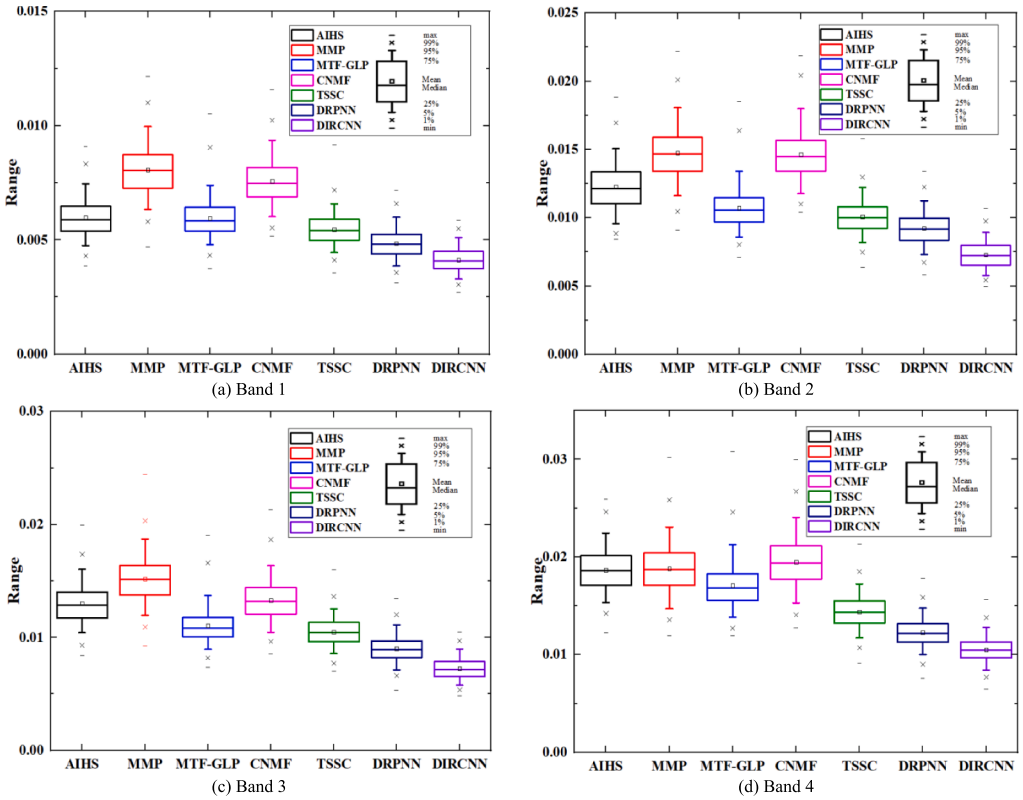

Fig. 10. Boxplots of the absolute difference between the different methods and the reference image of Fig. 9 in each band. Each box is represented as the absolute difference of a compared method and the reference image.

in Fig. 11(h), it can be seen that the bare soil area of the reference image is yellow, whereas that of CNMF is red; moreover, the spatial information in the vegetation area of CNMF does not appear to have increased. Thus, CNMF not only performs poorly in the spatial detail enhancement, but also causes obvious spectral distortion. The spatial information in the fusion result of TSSC is more than that in the results of the first four methods; however, some of the information is fake information, which does not exist in the reference image, as can be seen in the bare soil area of the zoomed area in Fig. 11(e). The fusion results of DRPNN and DIRCNN are closer to the reference image than the first five fusion results. Furthermore, the spatial details in the fusion result of DIRCNN are much clearer than in the result of DRPNN, and are the closest to the reference image, as can be seen in the vegetation part of the zoomed area in Fig. 11(f–h).

To further analyze the fusion results, Fig. 12(a–g) display the true-color synthesis (i.e. 3–2-1 band combination) of the absolute residual images between the seven fusion results and the reference image of Fig. 11, and Fig. 12(h) displays the absolute residual image between the Up-LR-MS image and the reference image of Fig. 11. In these residual images, the brighter the image, the greater the difference between the fusion result and the reference image. As shown in Fig. 12, the residual images of the seven fusion methods are darker than that of the Up-LR-MS image, which shows the effectiveness of these fusion methods. In addition, the residual images of the two CNN-based methods are darker than those of the five non-CNN-based methods, and in the two CNN-based methods, less spatial structure information exists in the DIRCNN residual image than in the DRPNN residual image.

4.5. Real-data experiments

Real-data experiments are performed on two data sets: IKONOS data sets and GF-1 data sets. When performing the real-data IKONOS

Table 6
Quantitative results of the simulated GF-2 images (20 groups)

Algorithm	Ideal data	AIHS	MMP	MTF-GLP	CNMF	TSSC	DRPNN	DIRCNN
ERGAS	**0**	2.9722	2.6631	2.8745	3.0972	2.7827	2.0402	1.8098
SAM	**0**	3.3881	3.3068	3.3310	3.5071	3.4694	2.6568	2.4412
Q	**1**	0.8937	0.9053	0.8986	0.8812	0.9133	0.9415	0.9557
PSNR	$+\infty$	33.8853	34.2007	33.1970	33.5423	33.9201	36.2941	36.9075
$SSIM_B$	**1**	0.9805	0.9845	0.9814	0.9749	0.9856	0.9880	0.9926
$SSIM_G$	**1**	0.9738	0.9883	0.9886	0.9770	0.9902	0.9894	0.9936
$SSIM_R$	**1**	0.9603	0.9812	0.9836	0.9712	0.9853	0.9850	0.9913
$SSIM_{NIR}$	**1**	0.8194	0.7996	0.7792	0.8230	0.8376	0.8931	0.9004
$SSIM_{AVG}$	**1**	0.9335	0.9384	0.9332	0.9365	0.9497	0.9639	0.9695
SCC_B	**1**	0.8288	0.8222	0.8150	0.7920	0.8030	0.9133	0.9323
SCC_G	**1**	0.8226	0.8243	0.8300	0.8012	0.8169	0.9101	0.9345
SCC_R	**1**	0.7776	0.7844	0.7940	0.7692	0.7831	0.8847	0.9131
SCC_{NIR}	**1**	0.4846	0.5268	0.4131	0.5032	0.3599	0.6402	0.6626
SCC_{AVG}	**1**	0.7284	0.7394	0.7130	0.6907	0.7164	0.8371	0.8606

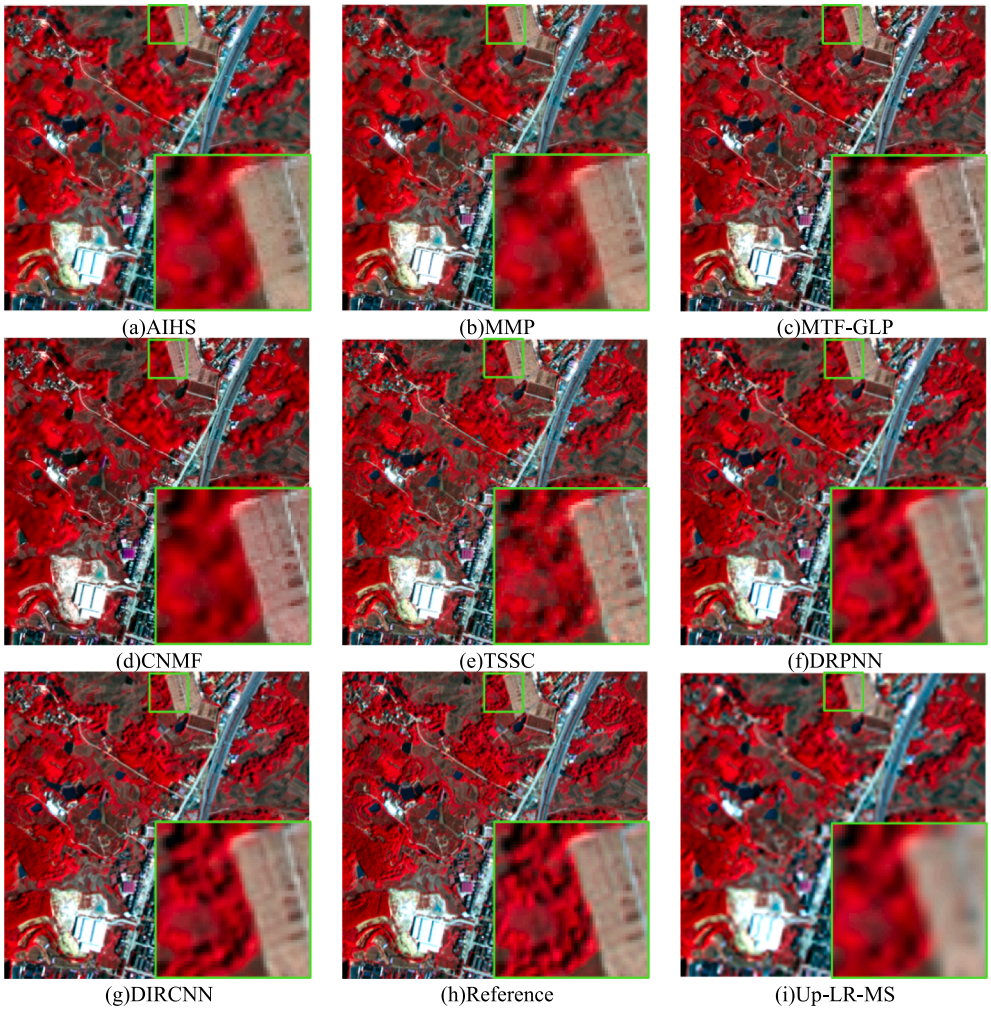

Fig. 11. Simulated GF-2 fusion results.

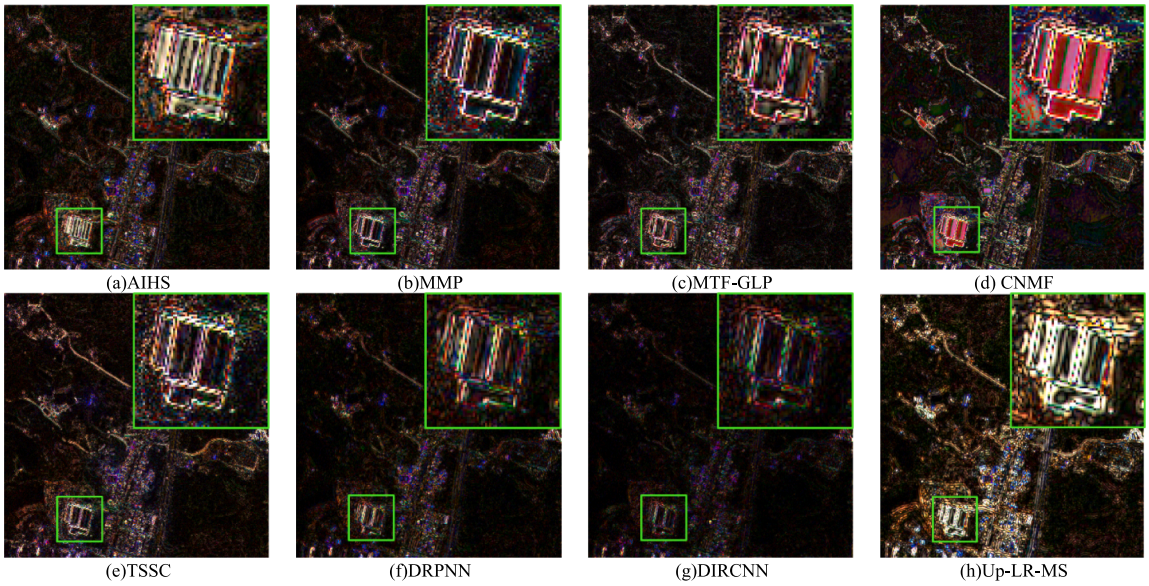

Fig. 12. Absolute residual images of the simulated GF-2 experiments.

Table 7
Quantitative results of the real-data ikonos images (four groups)

Algorithm	Ideal data	AIHS	MMP	MTF-GLP	CNMF	TSSC	DRPNN	DIRCNN
QNR	**1**	0.8383	0.8302	0.6918	0.8170	0.8132	0.9070	0.9153
D_λ	**0**	0.0412	0.0650	0.1351	0.0772	0.0699	0.0219	0.0192
D_S	**0**	0.1259	0.1144	0.2002	0.1153	0.1268	0.0729	0.0612
entropy	$+\infty$	6.7331	6.7546	6.8636	6.6515	6.7719	6.7114	6.7518
SF	$+\infty$	7.3092	11.9239	14.1653	10.8608	13.3355	6.4522	9.2755

experiments, the networks trained on the QuickBird images are utilized in all the CNN-based methods in the comparison. When performing the real-data GF-1 experiments, the networks trained on the GF-2 images are adopted in all the CNN-based methods.

4.5.1. Real-data IKONOS experiments

Table 7 lists the quantitative evaluation results of the real-data IKONOS experiments, with the average of four groups, where the best performance for each index is marked in red, the second-best performance is marked in blue, and the third-best performance is marked with underline. Among the different indices, QNR requires the HR-PAN image and the LR-MS image as auxiliaries. More specifically, D_λ is calculated by the fusion result and the LR-MS image, which shows the spectral quality; and D_S is calculated by the fusion result and the HR-PAN image, which shows the spatial quality. The entropy and SF values are obtained from the fusion results. The larger the entropy and SF values, the more spatial information the image has. As displayed in Table 7, the two CNN-based methods perform better in the QNR, D_λ, and D_S indices, which shows that the fusion results of the two CNN-based methods are more consistent with the LR-MS and HR-PAN images. MTF-GLP and TSSC perform better in the entropy and SF indices, which indicates that the fusion results of MTF-GLP and TSSC have more information than the other fusion results. When combined with the QNR, D_λ, and D_S indices, it can be speculated that some fake information exists in the fusion results of MTF-GLP and TSSC.

Fig. 13 displays the true-color synthesis (i.e. 3-2-1 band combination) of the various fusion images of a group of real-data IKONOS images. Compared with the Up-LR-MS image in Fig. 13(a), obvious spectral distortion appears in the fusion results of MTF-GLP and CNMF, as can be observed in the green vegetation area in the upper-left corner of Fig. 13(d) and the blue building area in the lower-right corner of Fig. 13(e). The spatial structure in the fusion result of AIHS is blurred, as can be seen in the building area in the lower-right corner of Fig. 13(b), and some tail trace exists in the vegetation area in the upper-left corner of the TSSC fusion image in Fig. 13(f), which is more obvious in Fig. 14(f). Overall, the fusion results of MMP, DRPNN, and DIRCNN are visually superior.

To further analyze these images, a small vegetation area and a small building area are selected to be zoomed in Fig. 14, where the zoomed vegetation area is displayed in the first row and the zoomed building area is shown in the second row. Note that, because the vegetation areas in the images in Fig. 13 are too dark, for a better comparative analysis, the brightness of the selected vegetation area in the first row is set to 50%. In the vegetation area row, compared with the Up-LR-MS image in Fig. 14(a), MMP, MTF-GLP, and CNMF show a weak spectral preservation performance, as can be seen in Fig. 14(c–e). Specifically, unlike the dark green vegetation in Fig. 14(a), the vegetation in Fig. 14(c) appears cyan, that in Fig. 14(d) appears gray-white, and that in Fig. 14(e) appears black. The most spatial information can be seen in Fig. 14(d), but some of this is fake information or noise. For example, the bare soil area surrounded by vegetation in Fig. 14(a–h) should be homogeneous, but there are many gray and white noise points in the result of MTF-GLP in Fig. 14(d). The TSSC image shows obvious tail trace in Fig. 14(f). DRPNN and DIRCNN perform well with regard to spectral fidelity, and the spatial structure in DIRCNN is a bit clearer than that in DRPNN, as shown in the vegetation next to the bare soil in Fig. 14(g–h).

In the building row, unlike the blue building in Fig. 14(i), the building in Fig. 14(l) appears blue-purple, and that in Fig. 14(m) appears dark blue, which shows the poor spectral preservation of MTF-GLP and CNMF. The fusion results of AIHS and DRPNN show blurred spatial information, as can be seen in the edge of the blue building in Fig. 14(j) and (o). Moreover, comparing the DIRCNN image in Fig. 14(p) with the MMP and TSSC images in Fig. 14(k) and (n), it can

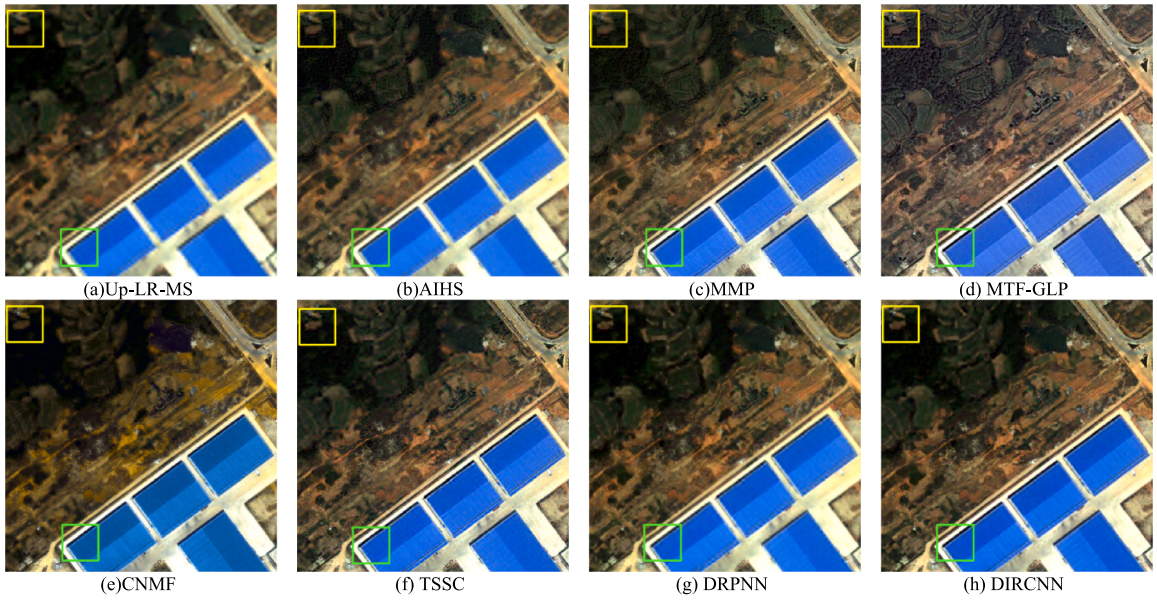

Fig. 13. Real-data IKONOS experiments.

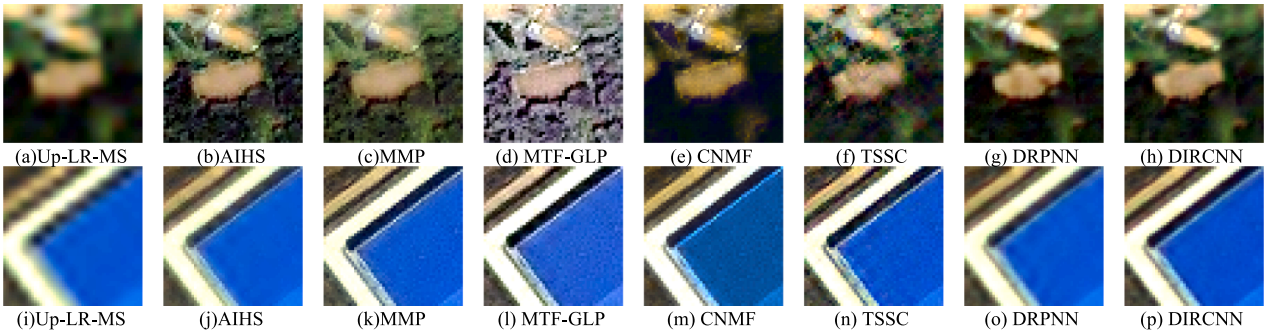

Fig. 14. The zoomed vegetation area and building area of the real-data IKONOS fusion images.

Table 8
Quantitative results of the real-data GF-1 images (four groups)

Algorithm	Ideal data	AIHS	MMP	MTF-GLP	CNMF	TSSC	DRPNN	DIRCNN
QNR	**1**	0.8068	0.8857	0.6999	0.7753	0.9175	0.9501	0.9390
D_λ	**0**	0.0586	0.0403	0.1502	0.1096	0.0364	0.0079	0.0283
D_S	**0**	0.1439	0.0772	0.1769	0.1300	0.0479	0.0424	0.0337
entropy	$+\infty$	6.6410	6.5369	6.7769	6.6589	6.6246	6.6672	6.6745
SF	$+\infty$	8.7210	5.8976	13.2122	12.4613	7.2526	6.1564	7.11328

be seen that the DIRCNN image seems like the denoising result of MMP and TSSC, which retains the spatial features while removing the noise information.

4.5.2. Real-data GF-1 experiments

The second real-data experiments are conducted on the GF-1 images. Table 8 lists the quantitative evaluation results of the real-data GF-1 experiments, with the average of four groups. As in the real-data IKONOS experiment results listed in Table 7, in the real-data GF-1 experiments, the two CNN-based methods perform better in the QNR, D_λ and D_S indices, and MTF-GLP performs best in the entropy and SF indices.

Fig. 15 displays a group of real-data GF-1 experimental results in true-color synthesis (i.e. 3-2-1 band combination). As displayed in Fig. 15, compared with the color of the Up-LR-MS image in Fig. 15(a), the fusion result of MMP appears to be covered with a pale green veil, and the vegetation in the fusion result of CNMF appears black, which shows the spectral distortion of MMP and CNMF. MTF-GLP also shows some spectral distortion in the vegetation area in Fig. 15(d). AIHS, TSSC, and DRPNN show blurred spatial information in Fig. 15(b) (f) (g), respectively, as shown in the building area of the lower-left corner of these fusion results.

For a more detailed comparison, small vegetation and building areas are selected to be zoomed in Fig. 16, where the first row displays the zoomed vegetation area in false-color synthesis, and the second row displays the zoomed building area in true-color synthesis. By comparing the zoomed vegetation images in the first row, it can be seen that there are varying degrees of visible spectral distortion in the AIHS, MMP, MTF-GLP, and CNMF images in Fig. 16(b-e), respectively. Specifically, referring to the Up-LR-MS image in Fig. 16(a), the results of AIHS, MMP, and MTF-GLP are whiter, and the result of CNMF appears a more vivid scarlet. The results of TSSC, DRPNN, and DIRCNN in Fig. 16(f–h) are closer to the LR-MS image in color, but TSSC and DIRCNN show a bit more spatial details than DRPNN. In the building images in the

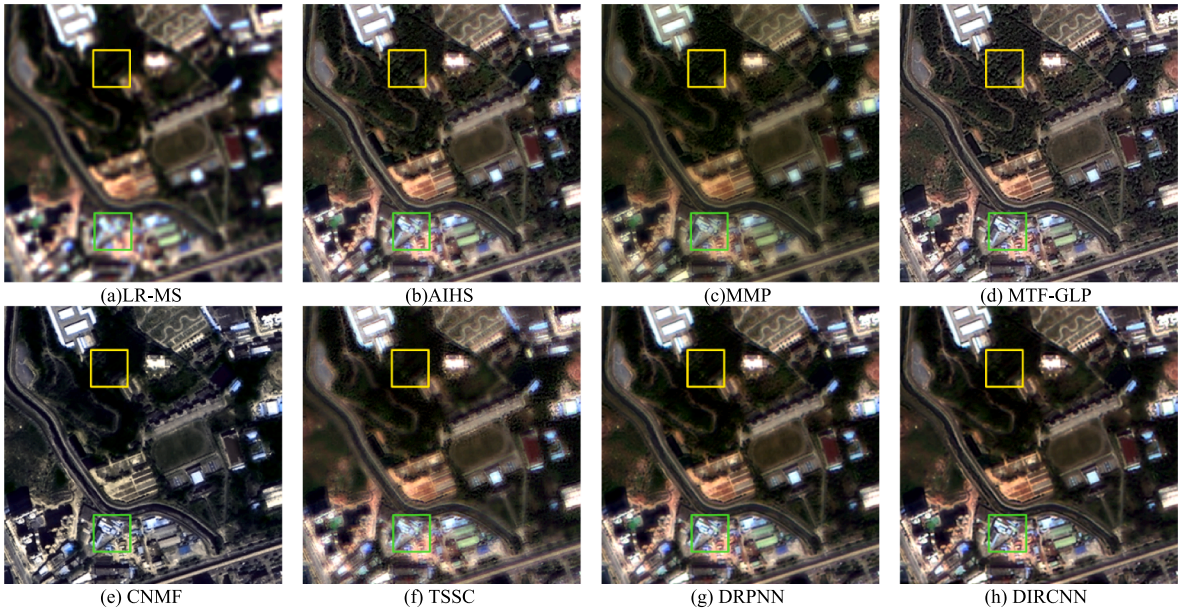

Fig. 15. Real-data GF-1 experiments.

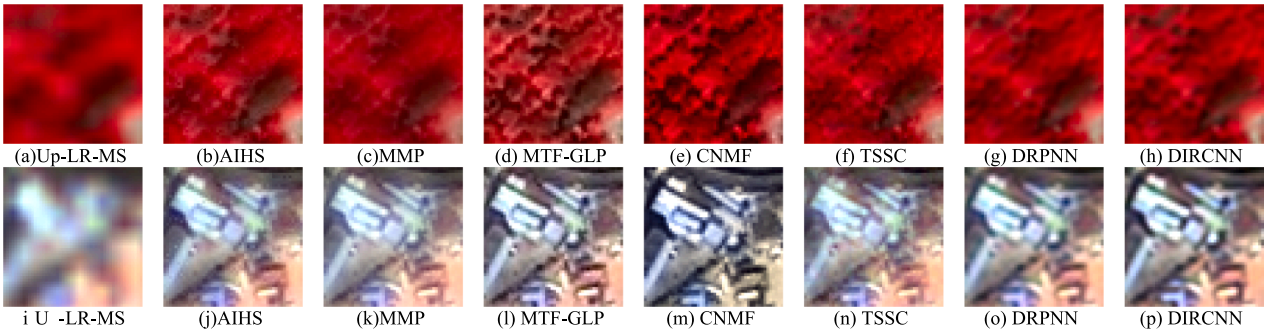

Fig. 16. The zoomed vegetation area and building area of the real-data GF-1 fusion images.

second row, CNMF shows very poor color contrast in Fig. 16(m), for example, the bare soil around the building should be orange, but that in Fig. 16(m) appears yellow. The AIHS, MMP, TSSC, and DRPNN images exhibit blurred spatial details, as can be seen in the edge of the building in Fig. 16(i) (k) (n), and (o), respectively. Overall, the results of MTF-GLP and DIRCNN are much more acceptable than those of the other approaches, as can be seen in Fig. 16(l) and (p).

4.6. Discussion

4.6.1. Parameters and complexity

For a more comprehensive analysis, the test time, training time, parameter number, and floating-point operations (FLOPs) (Xie et al., 2017) are discussed in this section. In Table 9, the first line lists the average test time of each method in each test data set, where the fastest time is marked in red, the second-best performance is marked in blue, and the third-best performance is marked with underline. It can be seen that the test time of two CNN-based methods are faster than that of other methods, of which DIRCNN is slightly slower than DRPNN. Among the five non-CNN based methods, the test time of AIHS and MTF-GLP is in the same order of magnitude, MMP takes slightly longer than AIHS and MTF-GLP, and two model-based methods take much longer. The second line lists the training time of DRPNN and DIRCNN, and the third line lists the parameter number of the fusion methods and the FLOPs of the two CNN-based methods. As indicated in Table 9, CNMF and TSSC require some manual parameter setting, which are set to achieve the satisfying fusion effect according to their original publications. In model-based methods, each parameter requires lots of manual experimental adjustments to find the optimal setting, the complexity of manual adjustment increases exponentially with the number of parameters. In contrast, DRPNN and DIRCNN learn their millions of parameters adaptively through the training process. In convolutional networks, in general, the more network parameters, the more complex the network, and the more time it takes for network training and testing. DIRCNN has a few more parameters and FLOPs than DRPNN, which makes DIRCNN more complex, with longer training and test times.

4.6.2. Advantages and disadvantages

As shown in Figs. 9–16, the proposed DIRCNN can make good use of the spatial information of the HR-PAN image to further improve the spatial information of the fusion results, while maintaining good spectral fidelity. Compared with DRPNN, the training and test times of DIRCNN are slightly longer, but the fusion accuracy is effectively improved.

Although DIRCNN can effectively improve the spatial details of the fusion image, the enhancement in vegetation area (especially continuous vegetation area) is not adequate. For example, as displayed in the zoomed areas in the lower-right corner of Fig. 11, compared with DRPNN in Fig. 11(f), DIRCNN in Fig. 11(g) shows more spatial details in the vegetation area, but still not as much as in the reference image in Fig. 11(h). To further improve the spatial details of the vegetation area, we will consider combining vegetation features, such as the normalized difference vegetation index (NDVI), in our follow-up work, to assist the training of the network.

5. Conclusions and future prospects

In this paper, a deep learning based pansharpening method that makes some unique changes to the input and the output of the fusion network is proposed. In more detail, the proposed method utilizes a novel differential information mapping strategy to assign the HR-PAN image to each band of the LR-MS image, to effectively improve the spatial quality of the fusion results. To further improve the accuracy, the gradient information of the Up-LR-MS image containing the rich spatial details of the ideal HR-MS image is utilized to assist the network. Moreover, the proposed network combines an attention module structure, which can recalibrate the extracted feature maps, and residual blocks, which can cascade the low-level features and the high-level features, to fully extract the features from the images. Experiments conducted on QuickBird, GF-2, IKONOS, and GF-1 images confirm the effectiveness of the proposed DIRCNN method, which can not only achieve a spectral fidelity that is as good as that of the other existing CNN-based approaches, but it also boosts the spatial details of the fusion images.

Table 9
Test time, training time, parameter number, and flops

	Algorithm	AIHS	MMP	MTF-GLP	CNMF	TSSC	DRPNN	DIRCNN
Test time(s)	QuickBird	0.2163	0.2766	0.2207	0.6503	3.0204	0.1120	0.1294
	GF-2	0.4834	0.7753	0.4821	1.7013	31.4304	0.3289	0.3758
	IKONOS	1.2320	2.8450	1.4258	7.2991	222.4818	1.0835	1.1969
	GF-1	1.2379	3.2164	1.4136	7.5626	141.4615	1.0842	1.1732
Training time	QuickBird	/	/	/	/	/	2h31min	3h44min
	GF-2	/	/	/	/	/	2h35min	3h41min
Parameters	/	/	/	/	6	4	1.64×10^6	1.66×10^6
FLOPs	/	/	/	/	/	/	1.57×10^9	1.58×10^9

The proposed DIRCNN method is found to work well on the PAN/MS (i.e. pansharpening) problem. In our future work, we will attempt to extend the method to the PAN/hyperspectral (HS) fusion problem and the MS/HS fusion problem. Furthermore, it should be noted that, in the real-data experiments of this paper, no networks is trained for the IKONOS images and the GF-1 images individually, as the network trained on the QuickBird images is used to fuse the IKONOS image, and the network trained on the GF-2 images is used to fuse the GF-1 images. The fusion results of DRPNN and the proposed DIRCNN for both the IKONOS images and the GF-1 images are acceptable, which indicates these networks' transferability between sensors. Due to the limited data available for some sensors, it will be of great significance to further explore the relationship between network structure and sensor transferability in the future.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under grant 61671334, grant 41701400, grant 41922008, and grant 61971319.

References

Aiazzi, B., Alparone, L., Baronti, S., Garzelli, A., Selva, M., 2006. MTF-tailored multiscale fusion of high-resolution MS and pan imagery. Photogramm. Eng. Remote Sens. 72 (5), 591–596.
Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B., 2006. A variational model for P+XS image fusion. Int. J. Comput. Vis. 69, 43–58.
Carper, W., Lillesand, T., Kiefer, R., 2004. The use of Intensity-Hue-Saturation transformations for merging spot panchromatic and multispectral image data. Photogramm. Eng. Remote Sens. 56 (4), 459–467.
Cheng, J., Liu, H., Liu, T., Wang, F., Li, H., 2015. Remote sensing image fusion via wavelet transform and sparse representation. ISPRS J. Photogramm. Remote Sens. 104, 158–173.
Duran, J., Buades, A., Coll, B., Sbert, C., Blanchet, G., 2017. A survey of pansharpening methods with a new band-decoupled variational model. ISPRS J. Photogramm. Remote Sens. 125, 78–105.
Fang, F., Li, F., Shen, C., Zhang, G., 2013. A variational approach for pan-sharpening. IEEE Trans. Image Process 22 (7), 2822–2834.
Fu, J., Liu, J., Tian, H., Fang, Z., Lu, H., 2018. Dual attention network for scene segmentation. arXiv preprint arXiv:1809.02983.
Ghahremani, M., Liu, Y., Yuen, P., Behera, A., 2019. Remote sensing image fusion via compressive sensing. ISPRS J. Photogramm. Remote Sens. 152, 34–48.
Gogineni, R., Chaturvedi, A., 2018. Sparsity inspired pan-sharpening technique using multi-scale learned dictionary. ISPRS J. Photogramm. Remote Sens. 146, 360–372.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. IEEE Conf. Comput. Vis. Pattern Recognit. 770–778.
He, L., Rao, Y., Li, J., Chanussot, J., Plaza, A., Zhu, J., Li, B., 2019. Pansharpening via detail injection based convolutional neural networks. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 12 (4), 1188–1204.
Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.
Jiang, C., Zhang, H., Shen, H., Zhang, L., 2014. Two-Step sparse coding for the pan-sharpening of remote sensing images. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 7 5, 1792–1805.
Kang, X., Li, S., Benediktsson, J.A., 2013. Pansharpening with matting model. IEEE Trans. Geosci. Remote Sens. 52 (8), 5088–5099.
Kiku, D., Monno, Y., Tanaka, M., Okutomi, M., 2013. Residual interpolation for color image demosaicking. IEEE Conf. Comput. Vis. Pattern Recognit. 2304–2308.
Kim, J.H., Choi, J.H., Cheon, M., Lee, J.S., 2018. RAM: Residual Attention Module for Single Image Super-Resolution. arXiv preprint arXiv:1811.12043.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Neural Inform. Process. Syst. 1097–1105.
Laben, C.A., Brower, B.V., 2000. Process for Enhancing the Spatial Resolution of Multispectral Imagery Using Pan-Sharpening. U.S. Patent 6011875.
Liu, X., Wang, Y., Liu, Q., 2018. PSGAN: a generative adversarial network for remote sensing image pan-sharpening. In: 2018 25th IEEE International Conference on Image Processing (ICIP) IEEE, pp. 873–877.
Martha, T.R., Kerle, N., Van Westen, C.J., Jetten, V., Kumar, K.V., 2012. Object-oriented analysis of multi-temporal panchromatic images for creation of historical landslide inventories. ISPRS J. Photogramm. Remote Sens. 67, 105–119.
Masi, G., Cozzolino, D., Verdoliva, L., Scarpa, G., 2016. Pansharpening by convolutional neural networks. Remote Sens. 8 (7), 594–615.
Meng, X., Shen, H., Li, H., Zhang, L., Fu, R., 2018. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis: practical discussion and challenges. Informat. Fusion 46, 102–113.
Moeller, M., Wittman, T., Bertozzi, A.L., 2008. Variational wavelet pan-sharpening. CAM Report 08–81.
Molina, R., Vega, M., Mateos, J., Katsaggelos, A.K., 2008. Variational posterior distribution approximation in Bayesian super resolution reconstruction of multispectral images. Appl. Comput. Harmonic Anal. 24, 251–267.
Nencini, F., Garzelli, A., Baronti, S., Alparone, L., 2007. Remote sensing image fusion using the curvelet transform. Informat. Fusion 8 (2), 143–156.
Palsson, F., Sveinsson, J.R., Ulfarsson, M.O., 2014. A new pansharpening algorithm based on total variation. IEEE Geosci. Remote Sens. Lett. 11 (1), 318–322.
Pohl, C., van Genderen, J.L., 1998. Multisensor image fusion in remote sensing: concepts, methods and applications. Int. J. Remote Sens. 19, 823–854.
Rahmani, S., Strait, M., Merkurjev, D., Moeller, M., Wittman, T., 2010. An adaptive IHS pan-sharpening method. IEEE Geosci. Remote Sens. Lett. 7 (4), 746–750.
Scarpa, G., Vitale, S., Cozzolino, D., 2018. Target-adaptive CNN-based pansharpening. IEEE Trans. Geosci. Remote Sens. 56 (9), 5443–5457.
Shahdoosti, H.R., Javaheri, N., 2017. Pansharpening of clustered MS and pan images considering mixed pixels. IEEE Geosci. Remote Sens. Lett. 14 (6), 826–830.
Shen, H., Jiang, M., Li, J., 2019. Spatial-spectral fusion by combining deep learning and variational model. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2019.2904659.
Shen, H., Li, T., Yuan, Q., Zhang, L., 2018. Estimating regional ground-level PM2.5 directly from satellite top of atmosphere reflectance using deep belief. Networks. J. Geophys. Res.-Atmos. 123 (24), 13875–13886.
Shen, H., Meng, X., Zhang, L., 2016. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. IEEE Trans. Geosci. Remote Sens. 54, 7135–7148.
Sirguey, P., Mathieu, R., Arnaud, Y., Khan, M.M., Chanussot, J., 2008. Improving MODIS spatial resolution for snow mapping using wavelet fusion and ARSIS concept. IEEE Geosci. Remote Sens. Lett. 5, 78–82.
Timofte, R., De Smet, V., Van Gool, L., 2014. A+: Adjusted anchored neighborhood regression for fast super-resolution. In: Asian Conference on Computer Vision. Springer, Cham, pp. 111–126.
Vivone, G., Alparone, L., Chanussot, J., Dalla Mura, M., Garzelli, A., Licciardi, G.A., 2014. A critical comparison among pansharpening algorithms. IEEE Trans. Geosci. Remote Sens. 53 (5), 2565–2586.
Wald, L., Ranchin, T., Mangolini, M., 1997. Fusion of satellite images of different spatial resolution: assessing the quality of resulting images. Photogramm. Eng. Remote Sensing 691–699.
Wang, S., Quan, D., Liang, X., Ning, M., Guo, Y., Jiao, L., 2018. A deep learning framework for remote sensing image registration. ISPRS J. Photogramm. Remote Sens. 145, 148–164.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.
Wei, Q., 2015. Bayesian Fusion of Multi-band Images: A Powerful Tool for Super-Resolution. Institut national polytechnique de Toulouse (INPT), 2015.
Wei, Y., Yuan, Q., Shen, H., Zhang, L., 2017. Boosting the accuracy of multispectral image pansharpening by learning a deep residual network. IEEE Geosci. Remote Sens. Lett. 14 (10), 1795–1799.
Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks. IEEE Conf. Comput. Vis. Pattern Recognit. 1492–1500.
Xing, Y., Wang, M., Yang, S., Jiao, L., 2018. Pan-sharpening via deep metric learning. ISPRS J. Photogramm. Remote Sens. 145 (A), 165–183.
Yokoya, N., Yairi, T., Iwasaki, A., 2011. Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion. IEEE Trans. Geosci. Remote Sens. 50 (2), 528–537.
Yuan, Q., Wei, Y., Meng, X., 2018. A multiscale and multidepth convolutional neural network for remote sensing imagery pan-sharpening. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 11 (3), 978–989.
Zhang, K., Zuo, W., Gu, S., 2017. Learning deep CNN denoiser prior for image restoration. IEEE Conf. Comput. Vis. Pattern Recognit. 3929–3938.
Zhang, L., Shen, H., Gong, W., Zhang, H., 2012. Adjustable model-based fusion method for multispectral and panchromatic images. IEEE Trans. Syst. Man Cybern. Part B: Cybern 42 (6), 1693–1704.
Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56 (8), 4274–4288.
Zhang, Y., Liu, C., Sun, M., Ou, Y., 2019. Pan-sharpening using an efficient bidirectional pyramid network. IEEE Trans. Geosci. Remote Sens. https://doi.org/10.1109/TGRS.2019.2900419.
Zhang, Y., Mishra, R.K., 2014. From UNB PanSharp to Fuze Go–the success behind the pan-sharpening algorithm. Int. J. Image Data Fusion 5 (1), 39–53.
Zhou, J., Civco, D.L., Silander, J.A., 1998. A wavelet transform method to merge Landsat TM and SPOT panchromatic data. Int. J. Rem. Sens. 19 (4), 743–757.