Research on super-resolution reconstruction of remote sensing images: a comprehensive review

Hui Liu,[a,b,c] **Yurong Qian,**[b,c,d,*] **Xiwu Zhong,**[b,c,d] **Long Chen**,[b,c,d] **and Guangqi Yang**[b,c,d]

[a]Xinjiang University, College of Information Science and Engineering, Urumqi, China
[b]Xinjiang University, Key Laboratory of Signal Detection and Processing, Urumqi, China
[c]Xinjiang University, Key Laboratory of Software Engineering, Urumqi, China
[d]Xinjiang University, College of Software, Urumqi, China

Abstract. The super-resolution (SR) reconstruction of remote sensing images is a low-cost and efficient method to improve their resolution, and it is often used for further image analysis. To understand the development of SR reconstruction of remote sensing images and research hotspots and trends, we examined its history and reviewed existing methods categorized into traditional, learning-based, and deep-learning-based methods. To evaluate the reconstruction performance, we conducted experiments comparing various algorithms for the single- and multi-frame SR reconstruction of remote sensing images considering three datasets. The experimental results indicate the advantages and limitations of single- and multi-frame reconstruction, with the latter showing a higher performance. Finally, we provide directions for future development of this SR reconstruction. [DOI: 10.1117/1.OE.60.10.100901]

Keywords: super-resolution reconstruction; remote sensing images; image analysis; methods categorized.

Paper 20210663V received Jul. 5, 2021; accepted for publication Oct. 11, 2021; published online Oct. 29, 2021.

1 Introduction

With the rapid development of image processing, numerous applications using digital images are emerging, with those using remote sensing images becoming a research hotspot. Remote sensing images provide wide coverage, rich information, and durability.[1,2] High-resolution (HR) remote sensing images are important in many fields, including environmental monitoring,[3] agricultural yield estimation,[4] urban planning,[5] military reconnaissance,[6] and emergency rescue.[7] However, owing to the high cost and long time required to develop HR remote sensing satellites, it remains challenging to conveniently obtain HR images. Super-resolution (SR) reconstruction has been devised to address this challenge by processing images to increase their resolution.[8,9] SR reconstruction is a traditional image processing problem. In recent years, many researchers have proposed SR technology to improve the spatial or spectral resolution of images.[10,11] We make an in-depth analysis on the status of SR reconstruction of remote sensing images from bibliometrics, such ScienceDirect, IEEE Xplore, Cnki Database, and Wanfang Database, as shown in Fig. 1. This method has low-cost and fast application and can be continuously improved, thus becoming practical and effective to expand the use of remote sensing images.[12] This review analyzes and discusses the field of SR reconstruction based on the research of domestic and foreign researchers. According to the number of input images, SR reconstruction method can be generally divided into two reconstruction schemes: single frame and multi-frame.[13] In the single-frame method, only a low-resolution (LR) image of the target scene can be used to generate an HR result, which was first proposed by Harris[14] and Goodman.[15] In the multi-frame method, multiple LR images of the same scene obtained under different conditions are used, first proposed by Tsai,[16] to improve the spatial resolution of Landsat TM images. In recent years,

[*]Address all correspondence to Yurong Qian, qyr@xju.edu.cn

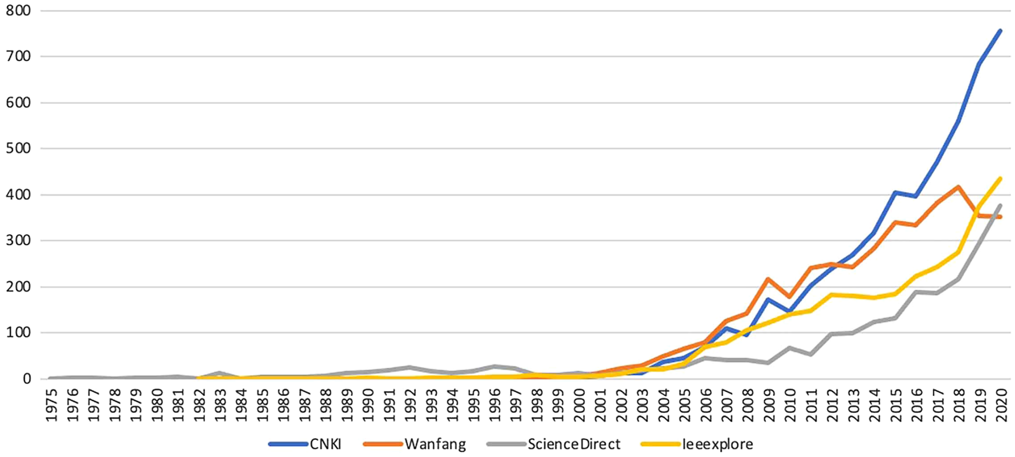

Fig. 1 SR reconstruction literature quantity trend chart.

the SR method of remote sensing image has been widely used. The most successful application is SPOT-5, which can generate an HR image with a resolution of about 2.5 m by processing two 5-m images.[17] With the increase of means to obtain images, multi-frame method has attracted more and more scholars' attention. In addition, with the development of deep learning, there are more and more layers of network, and the lightweight of network model has become the research direction of many scholars.[18]

Since the introduction of image SR reconstruction by Tsai[16] in 1984, this topic has been extensively studied, and there have been many in-depth discussions. In addition, SR reconstruction of remote sensing images has become a research hotspot. Moreover, machine learning and deep learning methods, especially deep neural networks, have shown remarkable reconstruction performance. In fact, machine learning can achieve a high performance by the following factors.

(1) Massive data can be used to train a deep neural network for tasks such as classification under the guidance of user-defined labels.

(2) The increasing capabilities of graphics processors allow processing and optimization in parallel.

(3) Deep neural networks can outperform traditional methods, and the model depth and output size of the image can be adjusted.

Although deep neural networks have contributed to remote sensing, various challenges remain to be addressed.

(1) Due to technical and budget constraints, a single-satellite sensor cannot acquire remote sensing images with high spatial and temporal resolution at the same time. Therefore, it is necessary to make effective use of time, space, and spectral correlation for multi-spectral and hyperspectral images from high latitudes.

(2) Remote sensing images mean long-distance observations encoded with a large number of sensors and scene information. Therefore, compared with natural images, remote sensing images contain richer information. Obtaining high-quality remote sensing images and extracting valuable characteristic information from them can provide a basis for remote sensing applications, such as remote sensing image classification applications.

(3) Remote sensing is affected by environmental factors, including atmospheric and weather conditions, which hinder the extraction of valuable information.

(4) Owing to the diversity in sampling methods, resolutions of remote sensing images, and available datasets, deep neural networks can be used to extract rich image information. Although deep learning can provide high performance, the computation burden increases with the complexity of the learning network, and the hardware requirements become prohibitive for practical large-scale applications.

This paper reviews the classification of remote sensing image SR reconstruction methods. Section 2 mainly discusses the traditional remote sensing image SR method. Section 3 mainly investigates the remote sensing image SR reconstruction method based on deep learning. Section 4 summarizes the quality evaluation criteria in remote sensing image reconstruction and the comparison between single-frame reconstruction and multi-frame reconstruction was done using some classical algorithms on three datasets. A summary and prospects are given in Sec. 5.

2 Based on Traditional Remote Sensing Image Super-Resolution Reconstruction

SR reconstruction is a traditional image processing problem, which mainly focuses on SR reconstruction for spatial resolution and spectral resolution. This review analyzes and discusses this field based on the research of domestic and foreign researchers on SR reconstruction. This section mainly summarizes the SR methods based on reconstruction and learning.

2.1 *Super-Resolution Reconstruction Method Based on Reconstruction*

SR reconstruction is a traditional SR problem. Considering the frequency domain, Tao et al.[19] used a discrete wavelet transform to decompose remote sensing images. Then the wavelet coefficient image was interpolated using nearest-neighbor, bilinear, or bicubic interpolation. The inverse discrete wavelet transform then provided the corresponding SR image. This method effectively preserves high-frequency information from the original HR image. To solve typical convolution and noisy linearly degraded images, Wei et al.[20] proposed a degraded image restoration algorithm based on a hidden Markov tree and Fourier-wavelet regularized deconvolution. Similarly, Jinliang et al.[21] used wavelet decomposition to obtain wavelet coefficients at different scales from high- and low-frequency information, obtaining weighting factors for suitable reconstruction, as demonstrated through multiple experiments. In addition, wavelet coefficients of multi-period LR images have been reconstructed using a wavelet transform to capture more details in the reconstructed images. Ma et al.[22] considered the frequency domain to reconstruct HR images in different frequency bands and proposed a method combining wavelet transform and the recursive ResNet architecture.

Among non-uniform image interpolation methods for reconstruction, the proposal by Tao et al.[19] combines the wavelet transform and an interpolation algorithm based on the image characteristics, improving the image resolution while preserving rich high-frequency information from the original remote sensing image. To prevent blurring of image boundaries after interpolation and loss of details, Jingmeng et al.[23] proposed an image SR reconstruction method based on the combination of residual pixels and non-uniform B-spline interpolation to reconstruct the visual effect, obtaining high-quality HR images. To use representative information from the medium- and high-frequency components in remote sensing images, Han[24] proposed an image interpolation algorithm based on the wavelet transform, which processes a bicubic interpolation image to preserve high-frequency information. Similarly, Solanki et al.[25] proposed a method to extract low- and high-frequency bands using a dual-tree complex wavelet transform and applying improved new edge-oriented interpolation to the high-frequency subband images, obtaining HR images without artifacts.

In the convex set projection method based on reconstruction, the traditional projection on convex sets (POCS) algorithm has contradictions in preserving image details and denoising, which affects the quality of the reconstructed image. In order to avoid this defect and obtain higher resolution, Shang[26] introduced the idea of image denoising based on sparse representation on the basis of POCS, combined the image processing method of sparse representation of K-singular value decomposition with the advantages of POCS for image SR reconstruction. Patti et al.[27,28] proposed another method of SR reconstruction of POCS, which took into account the blur factors caused by non-zero aperture time, camera movement, and imaging optical components.

The reconstruction-based maximum posteriori (MAP) method is more flexible, especially in the regular terms of the MAP method; one can freely add specific constraints on specific problems. For example, Tao et al.[29] introduced the Markov random field model on the basis of MAP to achieve SR reconstruction of sequence images. Markov random field theory imposes regularization constraints on LR to achieve rapid convergence and improve SR. Irmak et al.[30] proposed an improved method based on maximum posteriori-Markov random field (MAP-MRF) to enhance the spatial resolution of hyperspectral images. This method used non-linear programming technology to solve the joint energy minimization problem based on MAP-MRF, identified HR abundance maps, and combined them to obtain HR hyperspectral images which were very close to the original HR images. Irmak et al.[31] proposed a new MAP-based SR reconstruction method for hyperspectral images, in which the hyperspectral image was the only signal source. This method transformed the ill-posed SR reconstruction problem in the spectral domain into the quadratic optimization problem in the abundance mapping domain.

Iterative back projection based on reconstruction, conventional SR reconstruction based on iterative back projection causes a sawtooth effect and noise in the edge of reconstructed images, consequently compromising image clarity. To prevent these problems, Guo and Song[32] proposed image SR reconstruction based on a high-frequency enhancement curve and iterative back projection. The method first uses the unsharp mask to extract the high-frequency components of the initial reconstructed image, perform high-frequency information classification, and mitigate noise. Then the high-frequency enhancement curve is used to enhance the high-frequency components and maintain their monotonicity. As distortions caused by clouds, atmospheric turbulence, and other noise sources often vary across regions in LR remote sensing images, local distortions should be considered, but conventional iterative back projection cannot handle individual local distortions.[33,34] Thus Li et al.[35] improved iterative back projection by merging an inverse combination algorithm and a positive combination algorithm, improving elastic registration, and the image spatial resolution. Nevertheless, as iterative back projection is used for SR reconstruction of single-frame remote sensing images, the strong edges of the reconstructed image present the sawtooth effect. Tongyu[36] proposed an iterative process to handle image errors and further enhance the high-frequency components of images, thus improving the quality of image reconstruction and reflecting high-frequency information with stability and robustness (Table 1).

2.2 *Super-Resolution Reconstruction Method Based on Learning*

Learning-based methods have become important for SR algorithms in recent years.[28,30,37] Using a training set, such methods calculate the neighborhood relation between image blocks of test samples and training samples to determine the optimal weight constraint to obtain prior knowledge and approximate the HR image corresponding to the test sample. Therefore, sample-based SR algorithms based on neural network estimation have been proposed. Freeman et al.[38] introduced the application of sample-based methods for image SR reconstruction. Their algorithm uses a neural-network-based high-frequency small block estimation. Then a Markov network improves the resolution of the output small blocks. Wu and Wang[39] considered overfitting in traditional sample-based learning and regarded SR reconstruction as a regression problem. For a given set of training image pairs, they minimized the regularized cost function of the regressor to determine the minimum of the regressor in a sparse subset, reducing the time complexity and calculation cost while enriching image details such as textures.

Based on popular learning methods, the learning efficiency is not only related to the size of the training set but also to the use of the available samples. Su et al.[40] considered HR image blocks and their corresponding LR image blocks as points in high- and low-dimensional data spaces, respectively. Then they used local linear embedding to estimate LR small blocks that can map onto the corresponding HR small blocks. Specifically, the nearest neighbors of datapoints in the high-dimensional space from the local linear embedding were assumed to be the nearest neighbors in the low-dimensional space, and local linear embedding was used for SR. Chang et al.[41] proposed an SR algorithm based on neighborhood embeddings. This method uses a training set containing the corresponding LR and HR image blocks and assumes that the feature relation between any LR block and the training LR block is reconstructed to improve the resolution. Xinlei and Naifeng[42] proposed a sparse-structure manifold embedding method. By adding geometric rules of the image along the singular points or contours to neighborhood

Table 1 Summary of SR reconstruction methods based on reconstruction.

Algorithm	Advantage	Disadvantage
Frequency domain method	Image convolution, translation, rotation, and other operations can be easily converted into easy-to-handle arithmetic operations in the frequency domain	It is difficult to deal with the problem of image noise and can only deal with the spatially invariant noise model, and it is difficult to add prior information in the process. In addition, due to the complex transformation relationship between the frequency domain and the spatial domain, the traditional frequency domain method can only handle the situation where there is only global overall motion between the input LR images, and it is difficult to deal with the local motion, which has relatively large limitations
Non-uniform image interpolation	Very simple and intuitive, linking the problem of SR with the problem of image interpolation	The adaptability is relatively poor, it is difficult to deal with the blur phenomenon in the input image, the image introduces noise, etc., and it is difficult to add the prior information of the image
Convex set projection method	The thinking is relatively simple, the method form is relatively flexible, and the addition of prior knowledge is also relatively convenient	The computational complexity is high, and the convergence speed is relatively slow. In addition, its target solution is generally not unique but a set of feasible solutions
Maximum posterior probability	It has strong flexibility and robustness. Under the premise that the probability distribution of noise satisfies certain conditions, the original probability inference problem has a unique solution. At this time, you can choose an efficient gradient descent algorithm without worrying about converging to a local extreme	Need sufficient prior information
Example-based multi-linear regression method	Computational space complexity and time complexity are both low	The reconstruction quality depends on the anchor point scale and anchor point quality

selection, structure information of the image can be suitably restored. Simultaneously, by considering that abnormal values are often included in the embedding and that they reduce the structure accuracy, they used robust sparse embedding to eliminate outliers and normalized the weights to obtain more accurate neighborhoods and coding coefficients when synthesizing HR images. Zhang et al.[43] considered that multiple-point simulation based on linear dimensionality reduction compresses a high-dimensional space and shortened the simulation time, but it also reduces the simulation quality and limits the use for applications to non-linear data. Therefore, they introduced isometric mapping to achieve non-linear dimensionality reduction and combined multiple-point simulation with clustering for classification after dimensionality reduction. In addition, they designed a training process for image reconstruction to achieve continuous image feature reconstruction.

Based on the dictionary learning method, Yang et al.[9] proposed the use of sparse representation of image blocks to achieve SR reconstruction in 2008. An over-complete dictionary was formed by randomly selecting image blocks, and then linear programming was used for each test block. The method obtained the sparse representation of the test block under this over-complete dictionary and finally reconstructed an HR image with the weight of this set of coefficients. This method overcomed the problem of the neighborhood size selection in the neighborhood embedding method. Zhihui et al.[44] proposed a method based on sparse signal representation. This method jointly trained two dictionaries for LR and HR image blocks, respectively, and enhanced the similarity of sparse representation between them to realize the SR of remote sensing images and denoising. Liu et al.[45] proposed an image SR reconstruction algorithm based on sparse representation and classified texture patch, which mainly used prior knowledge and texture to reconstruct remote sensing images and performed dictionary learning by extracting image blocks from HR to LR. The super-complete dictionary was learned in the resolution image block, and then the trained dictionary was used to reconstruct the remote sensing image.

In the instance-based multi-linear regression method, it is considered that although the method based on dictionary learning has advantages in the quality of reconstructed images, it has a relatively high amount of calculation. Zhang et al.[46] proposed an instance-based support vector regression model, which learned the non-linear relationship between coarse fractional pixels and corresponding labeled subpixels from the selected best-matching training data. The coarse fractional images generated the HR land cover maps. It has more detailed spatial information and higher accuracy on different spatial scales. Subsequently, Zhang et al.[47] improved the method, which produced results with fewer spots and linear artifacts, more spatial details, smoother boundaries, and higher accuracy.

These methods either use the internal similarity in the image or learn the mapping between LR and HR image pairs. Although they focus on learning and dictionary optimization, the remaining steps of the methods are rarely optimized or considered under a unified framework.

3 Super-Resolution Reconstruction of Remote Sensing Image Based on Deep Learning

This section mainly summarizes the methods of SR reconstruction of remote sensing images based on deep learning models. Each type of model is divided into single-frame image and multi-frame image. Multi-frame SRR refers to the generation of a single high-quality and HR image from a group of low-quality and LR images (Table 2).

3.1 *Algorithms Based on BP Neural Network*

The back propagation (BP) neural network (its structure shows in Fig. 2) is a multi-layer feedforward neural network, which simulates the structure of the neural network of the human brain, and the basic unit of the human brain to transmit information is the neuron. There are a large number of neurons in the human brain, and each neuron connects with multiple neurons.[48] In the multi-frame remote sensing image SR reconstruction based on BP neural network, Ding et al.[49] used a three-layer BP neural network to reconstruct the resolution of the input remote sensing image. Because the BP neural network needs a lot of data to make it possible to converge, the

Table 2 Summary of learning-based SR reconstruction methods.

Algorithm	Advantage	Disadvantage
Example-based algorithm	Can learn *a priori* knowledge of various complex image structures and edges; can generate HR images with rich high-frequency information; and reduce the time complexity and calculation cost, and enrich the detailed information such as the texture of the image	The visual effect of the generated image is not good, contains noise, and the visual effect is improved but limited
Based on popular learning methods	Can make full use of the samples in the training set, and has good generalization ability under a small sample set	Sensitive to the selection of image features and the number of neighbors, which can easily lead to over-fitting or under-fitting
Dictionary-based learning method	It has the ability of adaptive field selection, low algorithm space complexity, and good robustness to noise	Reconstruction quality depends on dictionary size and dictionary quality
Example-based multi-linear regression method	Computational space complexity and time complexity are both low	The reconstruction quality depends on the anchor point scale and anchor point quality

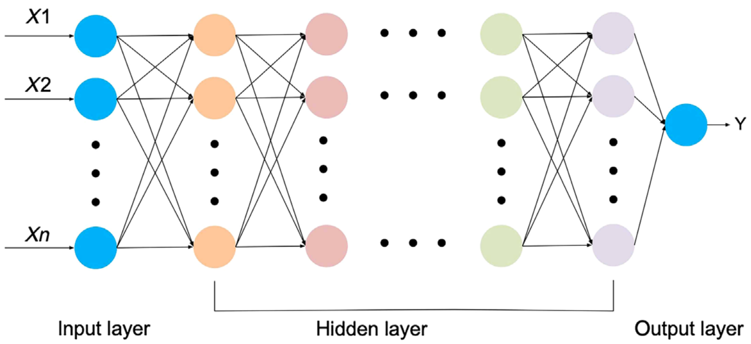

Fig. 2 BP neural network model diagram.

amount of calculations that it performs is also particularly prominent. Because the genetic algorithm can search for the global optimal solution, and it has strong robustness, but the convergence of the algorithm is deficient. Chen and Wang[50] combined genetic algorithm with BP neural network, so that the network has faster convergence ability and stronger learning ability.

3.2 *Algorithms Based on Convolutional Neural Network*

The development of convolutional neural networks (CNNs) can be traced back to the work by Hubel and Wiesel[51] on the visual system of the cat brain in 1962. Subsequently, Lecun et al.[52] proposed LeNet-5 in 1998. They introduced back projection to train neural networks, establishing the basis for CNNs. Nevertheless, it was not until 2012 when the AlexNet[53] was presented in the ImageNet large scale visual recognition challenge that CNNs began to flourish and to be widely used in various fields, achieving the highest performance at the time. A basic CNN consists of an input layer, a convolutional layer, a pooling layer, and a fully connected layer, as shown in Fig. 3. Moreover, the emergence of non-linear activation functions, such as the rectified linear unit (ReLU),[54] has increased the training performance of CNNs.

(1) Single-frame remote sensing image SR reconstruction based on CNN

Ducournau and Fablet[55] applied super-resolution CNN (SRCNN)[56] in handling ocean remote sensing data of large scale, the sea surface temperature field dataset, and obtained considerable gains in the peak-signal-to-noise ratio (PSNR) considering that in a typical CNN model, neurons in the lower convolutional layer share a smaller receptive field and pay more attention to local details, while in the higher layer, a larger receptive field is accumulated, covering a larger area. Lei et al.[57] proposed a novel image SR method called local global combination network (LGCNet). LGCNet has carefully designed the multi-fork structure, which can learn the multi-scale representation of remote sensing data, including local details (such as the edges and contours of objects) and global *a priori* (such as environment types). It is dedicated to reconstructing the residual between the LR and corresponding HR image pairs and achieved good robustness.

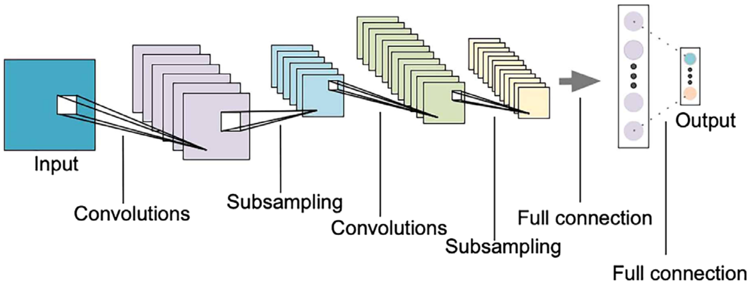

Fig. 3 Structure diagram of CNN.

(2) Multi-frame remote sensing image SR reconstruction based on CNN

Masi et al.[58] improved the SR-CNN and proposed a three-layer CNN for multi-spectral images with high spectral resolution and panchromatic images with high spatial resolution using feature fusion to generate multi-spectral images with high spatial resolution. Ye et al.[59] considered the spatial information in a multi-spectral image and used the available texture information to enhance the spatial resolution of a panchromatic image before processing, thereby improving the spatial resolution of the generated image. To better learn the characteristics of images across multiple frames, Palsson et al.[60] used a 3D CNN to fuse hyperspectral and panchromatic images to generate hyperspectral images with high spatial resolution. Yang et al.[61] used a two-branch CNN to extract spectral and spatial features of hyperspectral and multi-spectral images, respectively. The branch that processes hyperspectral images performs a 1D convolution, and its input is a 1D signal processed by a convolution. As a result, a remote sensing image with high spatial resolution and high spectral resolution is obtained.

3.3 *Algorithms Based on Generative Adversarial Network*

Generative adversarial networks (GAN) was first proposed by Goodfellow et al.,[62] and its structure is shown in Fig. 4. It is inspired by the zero-sum game in game theory and consists of a generator and a discriminator. The generator receives a random noise and generates a picture from the received noise. The discriminator is a binary classifier, which discriminates whether the input data are real data or the data generated by the generator. Both the generator and the discriminator can be implemented using any deep neural network model. Commonly used GAN network models include WGAN,[63] DCGAN,[64] InfoGAN,[65] EBGAN,[66] and LSGAN.[67]

(1) Single-frame remote sensing image SR reconstruction based on GAN

Ma et al.[22] proposed SR reconstruction for remote sensing images based on a transient GAN, which improved the previous SR-GAN.[68] Specifically, the conventional GAN was simplified by deleting components to reduce the memory requirements and increase the calculation speed. In addition, inspired by transfer learning, their reconstruction method was pretrained on the DIV2K dataset and then adjusted using a remote sensing image dataset, thus achieving high accuracy and visual performance. Similarly, Huang et al.[69] proposed an SR method based on GAN and residual learning for handling hyperspectral remote sensing images, extending the SR-CNN with a deeper architecture and more residual blocks. In addition, the generator was similar to that in SR-GAN, but the batch normalization layer was removed, and a deep network learned residual images. The combination of deeper layers and residual learning provided high spectral fidelity. Moreover, to improve the quality of spatial perception, a gradient learning network was developed to replace the VGG loss, and the discriminator was trained by transmitting the gradient features to the generator to contribute to detail recovery.

(2) Multi-frame remote sensing image SR reconstruction based on GAN

Liu et al.[70] were the first to develop a GAN for multi-frame image fusion using two-branch fusion as the generator of the required HR multi-spectral image and a fully convolutional network as the discriminator. Using this architecture, they obtained multi-spectral images with high

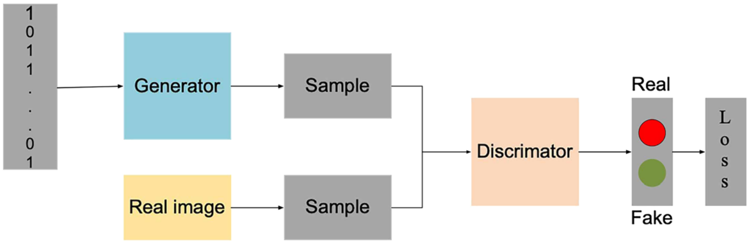

Fig. 4 GAN structure.

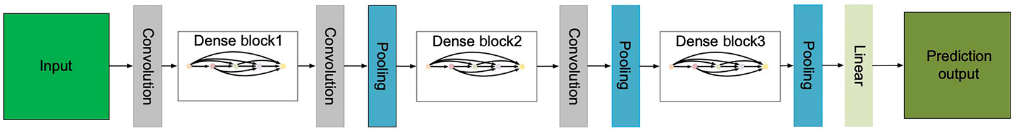

Fig. 5 DenseNet network structure.

spatial resolution. Shao et al.[71] further improved the results by introducing a residual encoder–decoder and a conventional GAN. Ma et al.[72] used a least squares GAN as the basic model and unsupervised learning. They adopted two discriminators, a spectral discriminator and a spatial discriminator, for the generated image to reflect both the spectral resolution of the multi-spectral image and the spatial resolution of the panchromatic image.

3.4 *Algorithms Based on Deep Dense Convolutional Network*

With the continuous improvement in deep learning networks, the problem of gradient messages has become persistent. To solve this problem, information transmission between adjacent layers in the network should be ensured. In addition, information transmission between the feature maps obtained in the early and late network stages should be considered. To maximize information transmission, Huang et al.[73] proposed a dense CNN, DenseNet, that transmits information across all the network layers. Specifically, each layer takes the results from all the previous layers as additional inputs. For example, for an L-layer network, a conventional CNN has L connections, whereas DenseNet has $L(L+1)/2$ connections. The DenseNet architecture is shown in Fig. 5. In the SR reconstruction of single-frame remote sensing image based on deep dense convolutional network, Pan et al.[74] proposed an single image super resolution (SISR) method based on residual dense back projection network to improve the resolution of an RGB image with medium and large-scale factors. The network is composed of dense back projection blocks, which contains two modules, called the upper projection module and the lower projection module. These modules are closely connected in one block to achieve better reconstruction performance, thereby making up for the fact that the details are ignored during the reconstruction process for large-scale factor.

3.5 *Algorithms Based on Deep Residual Network*

For single-frame SR reconstruction of remote sensing images based on deep dense CNNs, Pan et al. proposed single-image SR reconstruction based on the residual dense back projection network to improve the resolution of an RGB image with medium- and large-scale factors. Their network is composed of dense back projection blocks with an upper projection module and a lower projection module. These modules are closely connected in one block to improve reconstruction, compensating for details ignored during reconstruction for large-scale factors.

For conventional deep learning networks, it is generally believed that a deeper network with more parameters provides stronger non-linear expression ability and learning. However, the first problem caused by increasing depth is gradient explosion (or dissipation). This is because as the network has more layers, the gradient of BP in the network becomes unstable and excessively large or small as multiplications proceed. To prevent this problem, He et al.[75] proposed a residual module. For a stacked structure with several layers and input x and learned feature $H(x)$, residual $F(x) = H(x) - x$ should be learned for the original learning feature to be $H(x)$. When residual $F(x) = 0$, the accumulation layer only performs identity mapping, and the network performance is maintained. In practice, the residual cannot be zero, and the accumulation layer learns new features based on the input features to further improve the network performance.

(1) SR reconstruction of single-frame remote sensing image based on deep residual network

SR-GAN[68] introduces residual learning using multiple residual blocks in the generator (SR-ResNet). Each residual block contains two 3×3 convolutional layers followed by batch normalization and parametric ReLU activation. In addition, two subpixel convolutional layers are used to increase the feature size. The SR-GAN architecture is shown in Fig. 6.

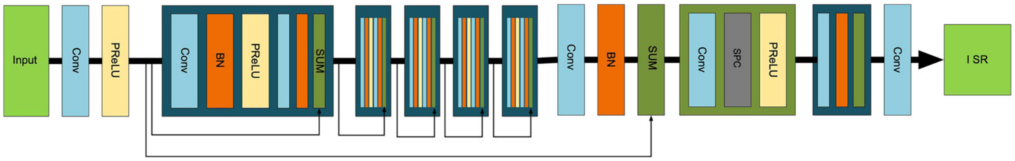

Fig. 6 SRGAN network model.

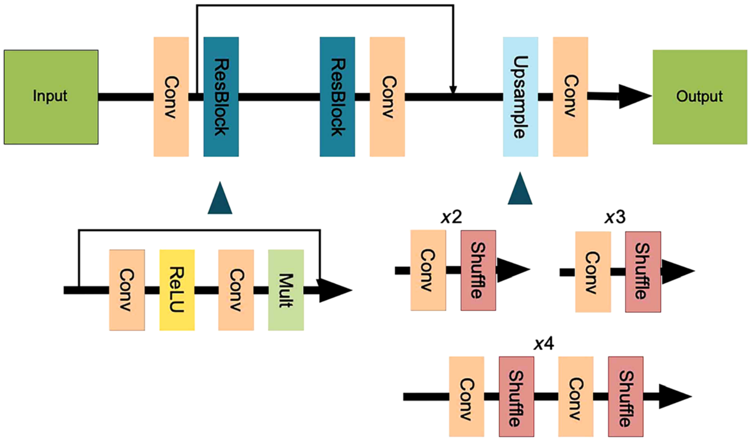

Fig. 7 EDSR network model structure.

Subsequently, enhanced deep residual networks[76] have considered the original ResNet intended to solve high-level computer vision problems, such as classification and detection. However, they do not directly apply ResNet to low-level computer vision problems, such as SR, because it provides suboptimal results. In addition, as batch normalization has the same memory requirements of the previous convolutional layer, enhanced deep residual networks remove this step (see Fig. 7), and instead it stacks more layers or improves feature extraction per layer to increase performance.

Wang et al.[77] improved an algorithm for SR reconstruction of remote sensing images based on a recursive residual network that combines global and local residual learning to facilitate deep network training under the control of recursive parameter learning. The improved recursive residual network uses global residual learning in the identity branch. In addition, recursive learning is introduced into the residual branch by constructing a recursive block stacked by residual units, and a multi-path structure is used in the recursive block. Ma et al.[22] combined the wavelet transform with a recursive ResNet. Thus frequency-domain reconstruction of HR images is achieved over different frequency bands. In detail, the wavelet transform is applied to LR images to divide them into various frequency components. Then a network with recursive residual blocks predicts high-frequency components. Finally, the image is reconstructed by applying the inverse wavelet transform.

(2) Multi-frame remote sensing image SR reconstruction based on deep residual network

In deep learning, high-frequency details are often lost due to the difficult learning of non-linear feature maps during fusion of multi-spectral and panchromatic images, and the spatial resolution of multi-spectral images should be improved. Therefore, Yang et al.[78] considered two residuals by gradually cascading them to learn non-linear feature maps from LR multi-spectral and panchromatic images to HR multi-spectral images. Zheng et al.[79] used a deep residual network to fuse hyperspectral and panchromatic images for the first time. In detail, they used a deep residual network to fuse edge-enhanced panchromatic images with initialized hyperspectral images that were generated by a guided filter to finally obtain remote sensing images with high spectral and spatial resolutions. To address the loss of details during upsampling and initialization of hyperspectral images and the restricted discriminative ability of CNNs caused by the equal treatment of various features, Zheng et al.[80] used a deep hyperspectral prior and a dual-attention residual network for multi-frame SR reconstruction with fusion of hyperspectral and panchromatic images.

3.6 *Algorithms Based on Feature Map Attention Mechanism*

In computer vision, an attention mechanism allows learning key features from essential information while ignoring irrelevant information. Depending on the application, attention mechanisms in image SR reconstruction perform channel attention or spatial attention.

(1) Single-frame remote sensing image SR reconstruction based on channel attention mechanism

The squeeze-and-excitation network (SENet)[81] considers that convolutions should improve the receptive field, that is, they should fuse features spatially or extract multi-scale spatial information. For feature fusion across channels, conventional convolutions basically fuse all channels of the input feature maps without considering the importance of different channels. Therefore, SENet considers the relationship between channels to adaptively recalibrate the feature response of individual channels by explicitly modeling interdependencies using a squeeze-and-excitation module, as shown in Fig. 8.

The SE module first squeezes the feature map obtained by convolution to obtain the channel-level global feature ($1 \times 1 \times C$), and then performs the excitation operation on the global feature, learns the relationship between each channel, and also obtains the weight of different channels, and finally multiplies the original feature map to get the final feature. Essentially, the SE module performs attention or selection operations in the channel dimension. This attention mechanism allows the model to pay more attention to the channel features of the most information, while suppressing those unimportant channel features.

Residual channel attention networks (RCAN)[82] introduced the channel attention mechanism to image SR. This method made two improvements on the basis of the EDSR model and proposed a residual in residual (RIR) structure to form a very deep network. It consists of several residual groups with long jump connections. Each residual group contains some residual blocks with short-hop connections. At the same time, RIR allows rich low-frequency information. Bypassing multiple hop connections, the main network is formed to focus on learning high-frequency information. Then a channel attention mechanism is proposed to adjust the channel characteristics adaptively by considering the interdependence between channels, and its essence is to combine the SE module and the residual.

Haut et al.[83] devised a CNN to handle the complexity of remote sensing images. The network uses residuals and skip connections in a very deep architecture to transmit information processed at different abstraction levels and alleviate data degradation. In addition, internal feature extraction implements a visual attention mechanism, which is integrated into the deep learning architecture for SR reconstruction of remote sensing images. Moreover, the channel focus domain responds according to the characteristics of the channel and reduces the computation burden associated with low-frequency information.

(2) Spatial attention mechanism

The convolutional block attention module considers[84] that attention should locate relevant information and improve the expression of the focus areas. Hence, to emphasize meaningful features in the spatial and channel dimensions, this module applies channel and spatial attention

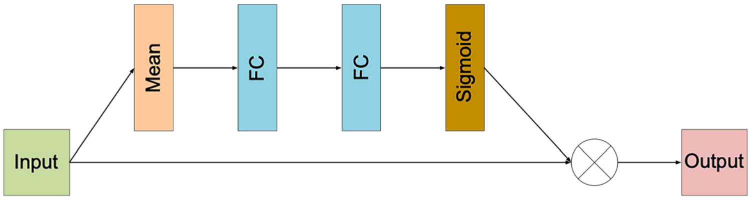

Fig. 8 Channel attention module.

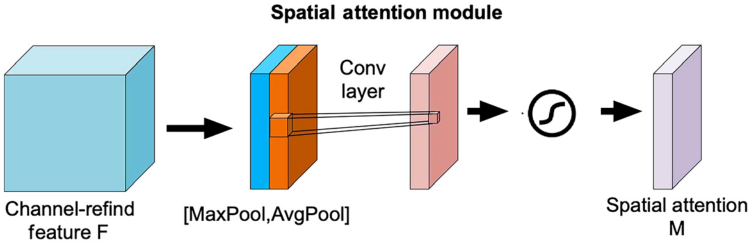

Fig. 9 Spatial attention module.

to learn salient areas and the attention results for the spatial and channel dimensions. The convolutional block attention module is basically the same as SENet for a channel, while the spatial dimension considers a spatial perspective, as shown in Fig. 9.

Similar to channel attention, given a feature F of $H \times W \times C$, it first performs average pooling and maximum pooling of a channel dimension to obtain two $H \times W \times 1$ channel feature maps, and these two feature maps are spliced together according to the channel. Then after a 7×7 convolutional layer, the activation function is Sigmoid, and the weight coefficient Ms is obtained. Finally, the weight coefficient and feature F are multiplied to get the new feature after scaling.

Residual feature aggregation (RFA) network[85] believes that the use of residual connections can improve network performance. Especially as the depth of the network increases, the residual features gradually concentrate on different aspects of the input image, which is very useful for reconstructing spatial details. However, the existing methods neglect to make full use of the hierarchical features on the residual branch. Therefore, an RFA framework is proposed for more effective feature extraction. The RFA framework groups several residual modules together and directly propagates features on each local residual branch by adding jump connections. At the same time, in order to maximize the function of the RFA framework, the improved enhanced spatial attention block is further used to make the residual features more focused on the key spatial content. Yang et al.[86] were inspired by U-Net and attention network, and added a spatial attention mechanism to the network of hyperspectral and multi-spectral image fusion. The spatial attention mechanism is used to retain more spatial information and generate the remote sensing images with high spatial resolution and high spectral resolution.

4 Experiment

4.1 *Quality Evaluation Criteria for Remote Sensing Image Reconstruction*

This paper uses eight widely used indicators to quantitatively evaluate the performance of the proposed method and the comparison methods.

The PSNR[87] reflects the quality of the reconstructed fused image by calculating the ratio of the maximum peak value of the reconstructed image to the mean squared error (MSE) of the two images. The PSNR is defined as

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right) = 20 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I}{\mathrm{MSE}}\right), \tag{1}$$

where MAX_I is the maximum value that represents the color of the image point. The higher the PSNR value is between two images, the less distorted the reconstructed image is relative to the HR image. The MSE is defined as

$$\mathrm{MSE} = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\|I(i,j) - K(i,j)\|^2, \tag{2}$$

where I and K are the two images of size $m \times n$, one of which is the noise approximation of the other.

The structural similarity (SSIM) index[88] measures the overall fusion quality by calculating the mean, variance, and covariance of the fused image and the reference image. The SSIM measurement consists of three contrast modules, namely, the brightness, contrast, and structure. Given two images X and Y of size $M \times N$, the means and variances of X and Y and the covariance of X and Y are represented by u_x, u_y, δ_x^2, δ_y^2, and δ_{xy}. The comparison functions that define the brightness, contrast, and structure are

$$l(X,Y) = \frac{2u_x u_y + c_1}{u_x^2 + u_y^2 + c_1}, \tag{3}$$

$$c(X,Y) = \frac{2\delta_x \delta_y + c_1}{\delta_x^2 + \delta_y^2 + c_1}, \tag{4}$$

$$s(X,Y) = \frac{\delta_{xy} + c_3}{u_x u_y + c_3}. \tag{5}$$

The combination of these three component factors is the SSIM indicator, which is defined as

$$\mathrm{SSIM}(X,Y) = [l(X,Y)]^{\alpha}[c(X,Y)]^{\beta}[s(X,Y)]^{\gamma}. \tag{6}$$

The closer the SSIM value is to 1, the higher the similarity is between the two images. The relative global dimensional synthesis error (ERGAS)[89] mainly evaluates the spectral quality of all the fusion bands within the spectral range, taking into account the overall situation of the spectral changes. It is defined as

$$\mathrm{ERGAS} = 100\frac{h}{l}\sqrt{\frac{\sum_{i=1}^{N}(\mathrm{RMSE}^2(B_i)/M_i^2)}{N}}, \tag{7}$$

where h is the resolution of the HR image, l is the resolution of the LR image, N is the number of bands, B_i is the MS image, and M_i is the average of the emissivity values of the MS image. The smaller the value is, the better the spectral quality of the fused image within the spectral range is.

The spectral angle mapper (SAM)[90] evaluates the spectral quality by calculating the angle between the corresponding pixels of the fused image and the reference image. It is defined as

$$\mathrm{SAM} = \arccos\left(\frac{(I_a J_a)}{\|I_a\|\|J_a\|}\right), \tag{8}$$

where I_α and J_α are the pixel vectors of the fused image and the reference image, respectively, at the distance point α. For an ideal fused image, the value of the SAM should be 0.

The spatial correlation coefficient (SCC) evaluates the similarity between the fused image and the spatial details of the reference image, uses a high-pass filter to extract the high-frequency information of the reference image, and calculates the correlation coefficient (CC) between the high-frequency information.[91] This paper uses a high Laplacian filter defined as

$$F = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix} \tag{9}$$

to obtain a high frequency. A higher SCC means that most of the spatial information of the PAN image is injected during the fusion process. The SCC is calculated between the fused image and the reference image. The final SCC is averaged over all bands of the MS image.

The CC is calculated as

$$\mathrm{CC} = \frac{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j} - \mu_X)(Y_{i,j} - \mu_Y)}{\sqrt{\sum_{i=1}^{w}\sum_{j=1}^{h}(X_{i,j} - \mu_X)^2(Y_{i,j} - \mu_Y)^2}}, \tag{10}$$

where X is the fused image, Y is the reference image, w and h are the width and height of the image, and μ represents the average value of the image.

The index Q[92] combines three factors to calculate image distortion: correlation loss, brightness distortion, and contrast distortion. It is defined as

$$Q = \frac{|\sigma_{Z1,Z2}|}{\sigma_{Z1} \cdot \sigma_{Z2}} \cdot \frac{2\sigma_{Z1} \cdot \sigma_{Z2}}{\sigma_{Z1}^2 + \sigma_{Z1}^2} \cdot \frac{2|\overline{Z_1}| \cdot |\overline{Z_2}|}{|\overline{Z_1}|^2 \cdot |\overline{Z_2}|^2}, \tag{11}$$

where Z_1 and Z_2 represent the b'th band of the fused image and the reference image. When Q is 1, it represents the best fidelity for reference.

The quality with no reference (QNR) index,[93] the original LRMS image, and PAN image are used to measure the direct spectral distortion of LRMS image and fused image, as well as the spatial distortion caused by the spatial detail differences generated by fusion.

The QNR is a non-reference image quality evaluation method. It is composed of the spectral distortion index D_λ and the spatial distortion index D_S:

$$D_\lambda = \sqrt[q]{\sum_{l=1}^{L} \sum_{\underbrace{r=1}_{r\neq 1}}^{L} \frac{|Q(I_l^{\text{HRMS}}, I_r^{\text{HRMS}}) - Q(I_l^{\text{LRMS}}, I_r^{\text{LRMS}})|^q}{L(L-1)}}, \tag{12}$$

$$D_S = \sqrt[q]{\frac{\sum_{l=1}^{L} |Q(I_l^{\text{HRMS}}, I_r^{\text{PAN}}) - Q(I_l^{\text{LRMS}}, I_r^{\text{LRPAN}})|^q}{L}}, \tag{13}$$

$$\text{QNR} = (1 - D_\lambda)^\alpha (1 - D_S)^\beta, \tag{14}$$

where the LRMS image with L bands is represented by I^{LRMS}, the generated HRMS image is I^{HRMS}, and only one band is the PAN image with I^{PAN}, and its degradation corresponds to the I^{LRPAN} image. The ideal value of the QNR index is 1, which means that the quality of the fused image is better.

With the development of single-image SR, the research on single-image SR is divided into two branches. One of them is based on PSNR and SSIM values, and the other is based on perceptual index (PI) values.[94] PI value represents the subjective perception quality of an image. Usually, the lower PI value is, the more comfortable the image looks. The lower the PI value is, the better the perceived quality of the image is, which is contrary to the PSNR value. In general, a low PI value is accompanied by a low PSNR value:

$$\text{PI} = \frac{1}{2}((10 - \text{Ma}) + \text{NIQE}). \tag{15}$$

Ma stands for Markov score and NIQE stands for natural image quality evaluator, that is, image evaluation quality.

4.2 *Experimental Analysis of Super-Resolution Reconstruction of Remote Sensing Images Based on a Single Frame*

This section conducts experimental comparison and data analysis on seven classic single-frame remote sensing image SR reconstruction models using three datasets.

4.2.1 *Introduction to dataset and experimental environment*

This section mainly introduces the experimental datasets and experimental environmental parameters for SR reconstruction of remote sensing images based on single frame.

(1) The datasets

We used three public datasets commonly used in the literatures related to remote sensing image SR, namely UCMerced_LandUse (UCM), aerial image dataset (AID), and PatternNet, to evaluate the classic SR methods.

UCM. This dataset was released by the University of California in 2010. According to the source is GoogleEarth, it includes 21 types of remote sensing scenes such as mediumresidential, airplanes, storagetanks, and parkinglot, each with 100 pictures. All images are of the size by 256×256 pixels, and the spatial resolution is 0.3 m/pixel. We randomly selected 40% of the images for training and 5% of the images for testing.

AID. This dataset was released by Wuhan University in 2012. The data source is GoogleEarth, including 30 types of remote sensing scenes such as parks, airports, mountains, and churches. Each type has 200 to 400 images, and all images are of the size by 600×600 pixels. The spatial resolution is 0.58 m/pixel. We randomly selected 100 images and 5 images in each class as the training dataset and the test dataset respectively.

PatternNet. This dataset was released by Wuhan University in 2018. The data source is GoogleMap, including 38 types of remote sensing scenes such as forest, freeway, railway, shipping_yard, and football field. Each category has 800 pictures. All images are of the size of 256×256 pixels, and the spatial resolution is 0.064.7 m/pixel.

(2) The experimental environment

We trained and tested the networks using three datasets, UCM, AID, and PatternNet. Each original image is regarded as an HR image and is down-sampled with a scale factor of 4 using the bicubic interpolation algorithm in the MATLAB environment as an LR image. The training sample is a set of 96×96 image patches randomly cropped from the HR image and the corresponding LR image. In the training process, the images for training were enhanced by random flips and rotations. We used the pytorch framework to implement all the experiments with the Adam optimizer. The initial learning rate was 5×10^{-4}, the minimum learning rate was 1×10^{-7}, the batch-size was 32, and a total of 500 epochs were performed (Tables 3 and 4).

4.2.2 *Quantitative comparison of different datasets*

The experiment compared seven classic algorithms, including the traditional interpolation-based algorithm (Bicubic), the CNN-based algorithm SRCNN,[68] the deep residual network-based algorithm SRResnet[68] and EDSR,[76] the dense residual network-based algorithm RDN,[95] as well as the algorithms RCAN[82] and DRN[96] with the attention mechanism (Table 5).

On the UCMerceed_LandUse dataset, due to the simple operation of bicubic interpolation, the convenience of calculation is the best in terms of time, calculation amount, and parameters.

Table 3 Summary of single frame image dataset.

Dataset	Scene type	Resolution (m/pixel)	Image size	Number of images in training sets	Number of images in test sets
UCM	21	0.3	256 × 256	840	105
AID	30	0.5 to 8	600 × 600	3000	150
PatternNet	38	0.06 to 4.7	256 × 256	3800	190

Table 4 Experimental environment configuration parameters.

Parameter	Numerical value	Parameter	Numerical value
Operating system	Windows 10	CUDA	CUDA11.0
CPU	i7-10700CPU at 2.90 GHz × 16	cudnn	cudnn-8.0
GPU	GeForce RTX 3070	Pytorch-GPU	1.7
RAM	32G/DDR4	GPU memory	8G

Table 5 Quantitative comparison of different algorithms under the condition of 4 times magnification on the UCM dataset. (Bold and italics indicate the best and the second best, respectively).

	Evaluation	PSNR	SSIM	PI	Flops	Params (M)	Time (s)
UCM	Bicubic	26.85	0.6978	7.7841	—	—	**0.001**
	SRCNN	27.82	0.7388	8.1992	**3.7**	**0.02**	*0.009*
	RDN	26.83	0.6961	9.7252	90.9	22.27	0.042
	SRResNet	29.09	0.7891	13.047	*9*	*1.52*	0.014
	EDSR	29.14	0.7901	9.3805	205.8	43.09	0.036
	RCAN	**29.26**	*0.7941*	12.7145	99.9	12.61	0.082
	DRN	*29.25*	**0.7945**	13.1988	94.7	4.80	0.038
	SRGAN_x4	25.55	0.6648	*4.6951*	*9*	*1.52*	0.014
	ESRGAN_SRx4	25.39	0.6685	**4.4418**	23	16.6	0.039

RCAN has achieved the best effect on PSNR, which is 2.41 dB higher than bicubic interpolation, and DRN is the best results that are achieved on SSIM, and the bicubic interpolation is 0.0517 higher. This is because compared with traditional algorithms; algorithms based on deep learning have better feature extraction capabilities and can better restore high-frequency information of images.

It can be seen from Table 6 that RCAN has achieved the best results in both PSNR and SSIM, and DRN is 0.02 and 0.009 lower than that in PSNR and SSIM. When zooming in 4 times, the test image size of UCM is 64×64, while the image size of the AID dataset is 150×150. Because of the different image input sizes, the calculation of FLOPS has also changed, and the calculation of EDSR the amount has increased by 5.5 times from 205 to 1130G, whereas SRCNN has increased by 7 times. The parameter amount of the model has not changed, which benefits from the parameter sharing of the deep learning model. The lower the PI index of the algorithm based on generating countermeasure network, the better the subjective visual effect of the image, but at the same time, the PSNR is also low.

On the PatternNet remote sensing image dataset, the RCAN remains the best-performer regarding the PSNR and SSIM measure. Considering that the size of its input image and UCM dataset are the same, the floating point operations (FLOPs) per second of the model do not

Table 6 Quantitative comparison of different algorithms under the condition of 4 times magnification on the AID dataset. (Bold and italics indicate the best and second best, respectively).

	Evaluation	PSNR	SSIM	PI	Flops	Params (M)	Time (s)
AID	Bicubic	27.98	0.6955	7.4847	—	—	**0.002**
	SRCNN	28.28	0.7105	7.1541	**14.9**	**0.02**	*0.011*
	RDN	28.81	0.7337	6.7693	363.8	22.27	0.058
	SRResNet	28.74	0.7305	6.8480	*36.1*	*1.52*	0.019
	EDSR	28.78	0.7322	6.8457	823.3	43.09	0.043
	RCAN	**28.89**	**0.7365**	6.8226	399.7	12.61	0.090
	DRN	*28.87*	*0.7356*	6.8771	379.1	4.80	0.047
	SRGAN_x4	26.15	0.6114	**2.6333**	*36.1*	*1.52*	0.019
	ESRGAN_SRx4	26.11	0.6270	*2.8914*	92.2	16.6	0.048

change, and the relative performance of other models is generally the same as that of UCM. Moreover, results on the AID dataset are consistent, showing that the deep learning-based method can achieve a high generalization ability. Thus the end-to-end RCAN is reliable and consistent across datasets.

4.2.3 *Qualitative comparison of different datasets*

The experiment compared seven classic algorithms, including the traditional interpolation-based algorithm (Bicubic),[19] the CNN-based algorithm SRCNN,[68] the deep residual network-based algorithm SRResnet,[68] EDSR,[76] the dense residual network-based algorithm RDN,[95] algorithms SRGAN[68] and ESRGAN[97] based on generating countermeasure network, as well as the algorithms RCAN[82] and DRN[96] with the attention mechanism. Tables 5–7 shows the quantitative analysis results of the seven algorithms based on three different datasets when the original image is enlarged 4 times. From these tables, it can be seen that there are three datasets.

(1) In terms of PSNR and SSIM, deep learning-based algorithms have achieved greater improvements in PSNR and SSIM compared to traditional interpolation algorithms (Bicubic). The algorithm based on the residual network (SRResnet) has achieved a 0.89-dB improvement on the UCM dataset compared with the algorithm based on the ordinary convolutional neural network (SRCNN), which confirms the effectiveness of the residual module in the SR field. EDSR has a larger network width than SRResnet and has achieved a 0.13-dB improvement on the UCM dataset. However, RCAN has reduced the width on the basis of EDSR and increased the depth of the model. It has achieved the best results on the PatternNet and AID datasets and obtained equivalent results if compared with DRN on the UCM dataset, which proves that a deeper model is better than a wider model.

(2) In terms of time, considering that the time taken by data reading, model loading, and file saving far exceeds the time taken by the data to be calculated on the model, the time calculation in this paper is to calculate time interval between the model reading data and the model producing the output which can better reflect the time difference between different models. The traditional method (Bicubic) only needs to perform direct interpolation due to its simple structure, which takes the shortest time. The time consumption of the algorithm based on deep learning has a great relationship with the number of layers of the model and the modules used. RCAN has a depth of 400 layers and takes 9 times longer than SRCNN with a three-layer structure. RDN takes into account the interaction between different layers and uses residual dense connections, so it takes 0.028 s more time on the UCM dataset than SRResnet, which takes 3 times more time. The depth of the

Table 7 Quantitative comparison of different algorithms under the condition of 4 times magnification on the PatternNet dataset. (Bold and italics indicate the best and second best, respectively).

	Evaluation	PSNR	SSIM	PI	Flops	Params (M)	Time (s)
PatternNet	Bicubic	26.46	0.6610	7.8897	—	—	**0.001**
	SRCNN	26.99	0.6887	7.9915	**3.7**	**0.02**	*0.009*
	RDN	27.99	0.6961	8.5969	90.9	22.27	0.042
	SRResNet	27.88	0.7254	8.5309	*9*	*1.52*	0.014
	EDSR	28.01	0.7306	8.4816	205.8	43.09	0.036
	RCAN	**28.07**	**0.7332**	8.5031	99.9	12.61	0.082
	DRN	*28.05*	*0.7317*	8.6296	94.7	4.80	0.038
	SRGAN_x4	25.77	0.6270	**4.2225**	*9*	*1.52*	0.014
	ESRGAN_SRx4	25.28	0.6140	*4.3635*	23	16.6	0.039

DRN model is not deep, but it uses the channel attention mechanism which needs more time.

(3) In terms of parameters and FLOPs, traditional algorithms do not have these two indicators due to their simple calculations. Only algorithms based on deep learning are counted and compared. It can be seen from the table that the model based on deep learning has gradually increased with the increase of evaluation indicators. The amount of parameters and calculations of the model are gradually increasing, especially FLOPs will increase exponentially when the input of test pictures becomes larger. EDSR has the largest width of the model, so its parameters and calculations are the most complex, reaching 1130G when the input image size is 150 × 150, but its PSNR and SSIM indicators are lower than that of RCAN and DRN, which once again proves that the impact of the depth of the model on performance is greater than the width of the model.

The qualitative comparison results of test images on the UCM dataset in Fig. 10 show that the images generated by bicubic interpolation are blurry and texture details are not reconstructed. Although the deep learning SR-CNN improves reconstruction, its results still show blurring. The more advanced SR-ResNet, DRN, and other networks achieve qualitative improvements in reconstruction by providing clear images and texture details. However, some areas are distorted, and the shapes and structures of small objects cannot be reconstructed using such methods.

The qualitative comparison results of test images from the AID dataset in Fig. 11 show that as the number of small objects in the image increases, the difficulty of reconstruction increases, and the algorithm based on bicubic interpolation misses most small objects. In contrast, deep learning networks show strong learning and feature extraction capabilities, achieving correct reconstruction. For instance, the white vehicle shows a clear comparison between the reconstructions of the evaluated algorithms. RCAN and DRN have achieved the best visual effects.

From Fig. 12, it can be seen that compared with the images on the UCM and AID datasets, the resolution of the image is lower, which increases the difficulty of reconstruction. The

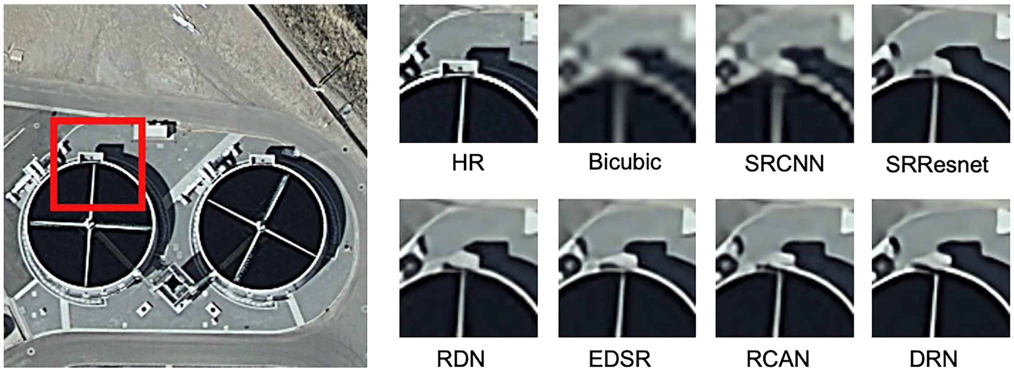

Fig. 10 Qualitative comparison results based on the UCM dataset at 4 times magnification.

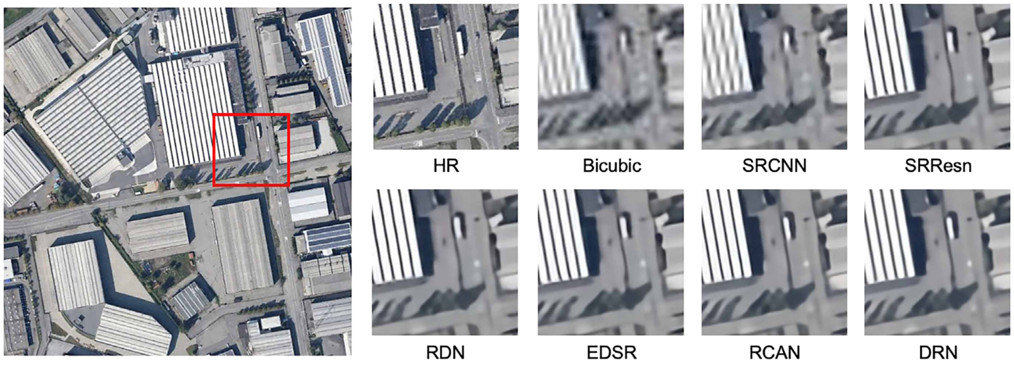

Fig. 11 Qualitative comparison results based on the fourfold magnification of the AID dataset.

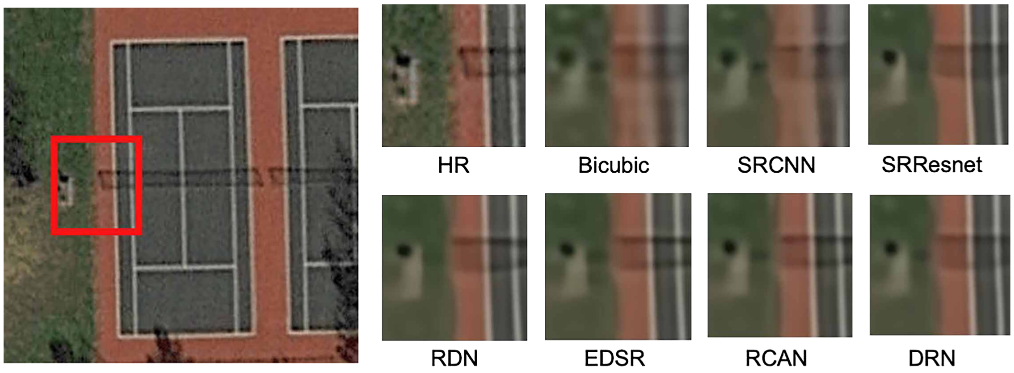

Fig. 12 Qualitative comparison results based on 4 times magnification on the PatternNet dataset.

interpolation algorithm and SRCNN cannot reconstruct the image at all, and the generated image effect is particularly poor. The image generated by SRResnet has some structural distortions. RDN, EDSR, RCAN, and DRN can better reconstruct high-quality images.

4.2.4 *Summary of comparative experiments*

In summary, the best algorithm is DRN, which has comparable results with RCAN in terms of PSNR and SSIM evaluation indicators, but it is better than RCAN in terms of parameter amount, calculation amount, and time. Other algorithms are superior to RCAN in terms of PSNR and SSIM. There are considerable advantages. The dual regression network adopted by DRN can reduce the number of model solution spaces, and the convergence speed will be faster. At the same time, the information of multi-scale images is effectively used in the upsampling process, which reduces the difficulty of reconstruction and achieves better visual effects. Although the two recent models, RCAN and DRN, have been separated for two years, they have not improved substantially. The DRN index of PSNR and SSIM is still lower, which indicates that the algorithm based on deep learning has encountered a bottleneck in the improvement of PSNR and SSIM index. In the future, it should be considered that on the premise of maintaining these two indices. A more lightweight and efficient model is developed to effectively reduce the number of model parameters and the amount of calculation.

4.3 *Experimental Analysis of Super-Resolution Reconstruction Based on Multi-Frame Remote Sensing Images*

This section conducts experimental comparison and data analysis on a variety of classic multi-frame remote sensing image SR reconstruction models using different datasets.

4.3.1 *Datasets and experimental environment*

This section conducts experimental comparison and data analysis under different datasets for a variety of classical multi-frame remote sensing image SR reconstruction models.

(1) The datasets

The datasets used in the experiment include the datasets from GeoEye-1, Spot-6, and GaoFen-2. The GeoEye-1 satellite is a commercial satellite launched by the United States from Vandenberg Air Force Base in California on September 6, 2008. It can acquire full-color images with a spatial resolution of 0.41 m and four-band (blue, green, red, and near-infrared) multi-spectral images with a spatial resolution of 1.65 m. The SPOT-6 satellite, launched on September 9, 2012, has a spatial resolution of 1.5 m for panchromatic images and 6 m for multi-spectral images, including blue, green, red, and near-infrared. The GaoFen-2 satellite, launched on August 19, 2014, has a spatial resolution better than 1 m. It is equipped with two cameras with an HR of 1 m panchromatic imaging and 4 m multi-spectral imaging. The relevant information of the three satellites GeoEye-1, Spot-6, and GaoFen-2 is shown in Table 8.

Table 8 Summary of multi-frame image dataset.

	Spectral wavelength (nm)					Spatial resolution (m)	
Satellite	PAN	Blue	Green	Red	NIR	PAN	MS
GeoEye-1	450 to 800	450 to 510	510 to 580	655 to 690	780 to 920	0.41	1.65
Spot-6	455 to 745	455 to 525	530 to 590	625 to 695	760 to 890	1.5	6
GaoFen-2	450 to 900	450 to 520	520 to 590	630 to 690	770 to 890	0.8	3.2

The three datasets used in this paper consist of three pairs of sizes of $320 \times 320 \times 4\text{-}1280 \times 1280 \times 1$, $256 \times 256 \times 4\text{-}1024 \times 1024 \times 1$, and $1000 \times 1000 \times 4\text{-}4000 \times 4000 \times 1$ multi-spectral-panchromatic image pair composition. The goal is to generate a multi-spectral image with the same size and the same spatial resolution as the PAN image. In order to evaluate the proposed model, we should compare the results obtained with non-existent reference images. According to the Wald protocol,[98,99] we use Gaussian blur to downsample the input image 4 times as the input to the network. The spectral image is used as a reference. We also use the bicubic interpolation algorithm to upsample the input spectral image to match the resolution of the PAN image. Here we use reference evaluation indicators: PSNR, SSIM, SAM, ERGAS, SCC, and Q.

In practical applications, there are no HR multi-spectral images for training. Therefore, we directly use the original data and the fusion reconstruction data to evaluate the indicators on the original resolution scale. Since there is no comparable ground truth, we use three non-reference image quality indicators D_Λ, D_S, and QNR.

(2) Experimental details

We train and test our network using three datasets, GeoEye-1, Spot-6, and GaoFen-2. In the training phase, we cut the upsampled MS image and PAN image into 32×32 image pairs and then randomly select 90% and 10% of the cut images as the training set and the validation set. In the testing phase, we then cropped the up-sampled MS image and PAN image to a 320×320 image pair, then filled the edges to 400×400 and input them to the network, and then cropped the image edges of the output network to restore to 320×320, then stitch all the test images to form a new generated image. The experiments were done using Keras and the configuration of the experimental environment is shown in Table 9. We used Adam optimizer to minimize losses. The initial learning rate lr is set to 0.0005. After 5 epochs, if the loss does not decrease, the learning rate is adjusted to $lr = 0.2 \times lr$. We performed a total of 600 epochs, and the batch size was set to 32.

4.3.2 *Quantitative comparison of different datasets*

We compared 15 algorithms in total, including bicubic-based SISR algorithm, Brovey-based transform,[100] PCA,[98] IHS-based component replacement,[99] SFIM-based brightness smoothing filter adjustment,[101] GS-based Gram–Schmidt transform,[102] wavelet-based transform,[103] generalized

Table 9 Experimental environment configuration parameters.

Parameter	Numerical value	Parameter	Numerical value
Operating system	Ubuntu 18.04	CUDA	CUDA11.0D
CPU	i7-10700CPU at 2.90 GHz × 16	cudnn	cudnn-8.0
GPU	GeForce RTX 3070	tensorflow-GPU	1.15.4
RAM	32G/DDR4	Keras-GPU	Keras 2.3.1

Table 10 Quantitative evaluation results on the GeoEye-1 dataset. (Bold, bold-italics, and italics, respectively, indicate the best, second best, and third best results).

	Reference comparison						No reference comparison		
Metrics	PSNR	SSIM	SAM	ERGAS	SCC	Q	D_λ	Ds	QNR
Bicubic	26.7239	0.6751	0.0867	4.8367	0.8858	0.4254	***0.0323***	***0.0274***	***0.9412***
Brovey	27.0411	0.7554	0.0934	4.0589	0.9183	0.593	0.1002	0.1063	0.8041
PCA	17.1862	0.6073	0.3942	12.484	0.6486	0.4028	0.0595	0.1687	0.7818
IHS	27.6877	0.7706	0.0888	3.8437	0.9272	0.6139	0.0897	0.1143	0.8063
SFIM	25.0983	0.7095	0.1004	5.8498	0.8481	0.5476	0.145	0.0529	0.8098
GS	27.8395	0.7791	0.0885	3.7992	0.9281	0.6251	0.0523	0.1022	0.8508
Wavelet	26.1692	0.6831	0.0951	5.0181	0.8732	0.4642	0.0448	0.054	0.9036
MTF_GLP	26.5291	0.7472	0.0937	3.8994	0.9139	0.6146	0.1842	0.0736	0.7557
MTF_GLP_HPM	25.1797	0.742	0.0977	5.3437	0.8801	0.6089	0.1913	0.0552	0.764
GSA	28.1781	0.7825	0.0851	3.591	0.9312	0.637	0.0624	0.0706	0.8714
CNMF	*28.5319*	*0.7851*	*0.0827*	*3.5742*	*0.9346*	*0.623*	0.061	0.0758	0.8678
GFPCA	26.1225	0.7031	0.1069	4.7886	0.8903	0.4259	0.0649	0.1257	0.8176
PNN	28.1373	0.7417	0.0831	4.0768	0.9187	0.5649	**0.0141**	**0.0163**	**0.9698**
PanNet	**30.497**	**0.8273**	**0.0657**	**2.8207**	**0.9593**	**0.7262**	*0.0364*	0.0313	0.9334
ResTFNet	***29.763***	***0.8213***	***0.0708***	***3.1999***	***0.9497***	***0.7135***	0.0365	*0.031*	*0.9336*

Laplacian pyramid MTF_GLP[104] of modulation transfer function, generalized puller with matched filter and multiple compound injection model Plass Pyramid MTF_GLP_HPM,[105] adaptive GS GSA,[106] CNMF-based coupled non-negative matrix factorization,[107] GFPCA-based PCA and guided filter[108] and deep learning-based PNN,[58] PanNet,[109] and ResTFNet.[110] The experimental quantitative evaluation results on the GeoEye-1 dataset are shown in Table 10.

It can be seen from Table 10 that in terms of reference evaluation indicators, PanNet and RestFNet have achieved the best and the second best results. This shows that deep learning methods are superior in spatial and spectral reconstruction to other algorithms. For other algorithms, such as CNMF based on coupled non-negative matrix factorization, good results have also been obtained.

In terms of non-reference measures, PNN achieves the best results, and ResTFNet improves its results. Overall, deep learning methods outperform other algorithms. The ResTFNet based on the two-branch fusion network can achieve better results with reference and no reference indicators. This indicates that the two-branch fusion network in the deep learning method is better than the method of directly stacking the input images in the processing of multi-frame remote sensing image superpartition reconstruction. It can use two subnetworks to process the two input images, respectively, and extract more effective features, so as to improve the quality of the fusion image.

As the consistent quantitative results of the GeoEye-1 dataset, the quantitative results of the Spot-6 dataset in Table 11 show that in terms of reference evaluation measures, the deep learning PENNet and ResTFNet also achieve the highest performance. Other algorithms such as CNMF also obtain suitable results. Again, ResTFNet achieves suitable results on non-reference evaluation measures, and other methods such as GSA also achieve good results. ResTFNet achieves high performance regarding reference and non-reference measures, indicating that deep learning is benefited from two-branch fusion, which outperforms stacking inputs in terms of spatial and spectral reconstruction. Overall, methods based on deep learning are superior to other algorithms for reconstruction.

Table 11 Quantitative evaluation results on the Spot-6 dataset. (Bold, bold-italics, and italics, respectively, indicate the best, second best, and third best results).

	Reference comparison						No reference comparison		
Metrics	PSNR	SSIM	SAM	ERGAS	SCC	Q	D_λ	Ds	QNR
Bicubic	23.8359	0.5484	0.1025	6.7785	0.8262	0.3961	**0.0236**	0.0569	**0.9209**
Brovey	24.3703	0.7417	0.1102	5.4564	0.8458	0.6695	0.0464	0.1065	0.852
PCA	20.3483	0.7066	0.3467	8.67	0.8829	0.6605	0.0935	0.0778	0.8359
IHS	25.1374	0.7432	0.1104	5.3214	0.87	0.6799	0.069	0.1021	0.8359
SFIM	23.9448	0.678	0.1033	6.4248	0.8337	0.5869	0.0954	0.0654	0.8455
GS	25.2396	0.7405	0.1124	5.3661	0.8746	0.6781	*0.0406*	0.0967	0.8666
Wavelet	23.7197	0.5937	0.1096	6.7671	0.8188	0.4587	***0.0352***	0.0753	0.8921
MTF_GLP	25.87	0.7713	0.094	4.6184	0.8941	0.7289	0.1242	0.0674	0.8168
MTF_GLP_HPM	25.7115	0.775	*0.0894*	4.7169	0.8912	*0.7306*	0.1195	0.0653	0.823
GSA	25.5626	0.7467	0.1031	5.0547	0.8843	0.6995	0.0824	0.0655	0.8574
CNMF	*26.3792*	*0.7795*	***0.0862***	*4.5675*	*0.9026*	0.7177	0.0883	0.0563	0.8604
GFPCA	24.1668	0.6599	0.1144	6.2197	0.8483	0.5174	0.0813	0.1277	0.8013
PNN	26.1832	0.7413	0.0954	4.7867	0.9023	0.6884	0.0625	**0.0305**	0.9089
PanNet	**27.9401**	**0.82**	**0.0862**	**3.9482**	**0.9366**	**0.7875**	0.0492	*0.0422*	*0.9107*
ResTFNet	***27.5324***	***0.8028***	0.0942	***4.137***	***0.9299***	***0.7725***	0.0432	***0.0378***	***0.9207***

The quantitative evaluation results of the Gaofen-2 dataset are listed in Table 12. As in Tables 10 and 11, PanNet and ResTFNet achieve the best results in terms of the reference evaluation measures. In addition, PNN achieves suitable results on this dataset. Regarding non-reference measures, RestFNet achieves suitable results. Overall, deep learning methods outperform the other algorithms on all the evaluated datasets. Remarkably, ResTFNet based on two-branch fusion achieves high performance regarding all the measures.

4.3.3 *Qualitative comparison of different datasets*

The qualitative evaluation results of the GeoEye-1 dataset are shown in Fig. 13. The images generated by PanNet and RestTFNet are visually superior to those generated by the other algorithms. Although the IHS method provides a suitable spatial reconstruction, it exhibits spectral distortion. The images obtained using bicubic interpolation for single-image SR reconstruction, wavelet transform, GFPCA, and guided filter show blur and artifacts. The first PNN model based on deep learning is also ambiguous. At the same time, MTF_GLP, MTF_GLP_HPM, SFIM, and PCA all have severe spectral and spatial distortions.

The qualitative evaluation results of the Spot-6 dataset are shown in Fig. 14 PanNet and RestFNet are visually superior to the other algorithms. The images obtained using bicubic interpolation, wavelet transform, and GFPCA still show blur and artifacts, while the spectral distortion of the IHS method is still present, and MTF_GLP, MTF_GLP_HPM, SFIM, and PCA show serious spectral and spatial distortions. Moreover, the ambiguity of PNN persists.

The qualitative evaluation results on the Gaofen-2 dataset are shown in Fig. 15. Seeing from Fig. 15, it indicated that PanNet and PNN are visually superior to other algorithms. The images obtained by Bicubic, wavelet, GFPCA, and CNMF methods have blur and artifacts. The spectral distortion of IHS still exists. MTF_GLP, MTF_GLP_HPM, SFIM, and PCA also have serious spectral and spatial distortions.

Table 12 Quantitative evaluation results on the Gaofen-2 dataset. (Bold, bold-italics, and italics, respectively, indicate the best, the second best, and the third best results).

	Reference comparison						No reference comparison		
Metrics	PSNR	SSIM	SAM	ERGAS	SCC	Q	D_λ	Ds	QNR
Bicubic	23.6379	0.5207	0.1333	9.7288	0.7972	0.3724	**0.0248**	0.2904	0.6921
Brovey	25.2994	0.799	0.1451	7.9374	0.8914	0.745	*0.0606*	0.1485	0.7999
PCA	23.9662	0.7284	0.2716	9.2124	0.8364	0.666	0.096	0.2286	0.6973
IHS	25.4087	0.7768	0.1796	7.8865	0.8831	0.7236	0.1047	0.1564	0.7553
SFIM	24.1771	0.7132	0.1865	9.0287	0.8393	0.6446	0.0824	***0.0532***	**0.8688**
GS	25.0856	0.7585	0.2066	8.1032	0.8713	0.7004	0.0744	0.1968	0.7435
Wavelet	24.0319	0.6628	0.1779	9.2471	0.8252	0.6036	0.1511	**0.0319**	0.8218
MTF_GLP	27.3228	0.8021	0.1932	6.2051	0.9271	0.763	0.1195	0.1165	0.7779
MTF_GLP_HPM	26.8807	0.8148	0.1993	6.4427	0.928	0.7717	0.0734	0.0889	***0.8442***
GSA	26.4543	0.7761	0.212	6.885	0.9189	0.7467	0.1604	0.1374	0.7243
CNMF	27.8433	0.8265	0.1567	5.9334	0.9291	0.7693	0.1048	0.0887	0.8158
GFPCA	23.6819	0.5894	0.1785	9.6273	0.83	0.397	***0.0352***	0.2126	0.7597
PNN	***29.9057***	**0.8677**	***0.127***	***4.689***	**0.9616**	**0.8472**	0.1011	0.0979	0.8109
PanNet	**29.9086**	***0.8625***	**0.1265**	**4.6767**	***0.9607***	***0.8421***	0.0891	0.0934	0.8259
ResTFNet	*29.2961*	*0.8578*	*0.1286*	*4.9978*	*0.9559*	*0.8354*	0.0894	*0.0884*	*0.8301*

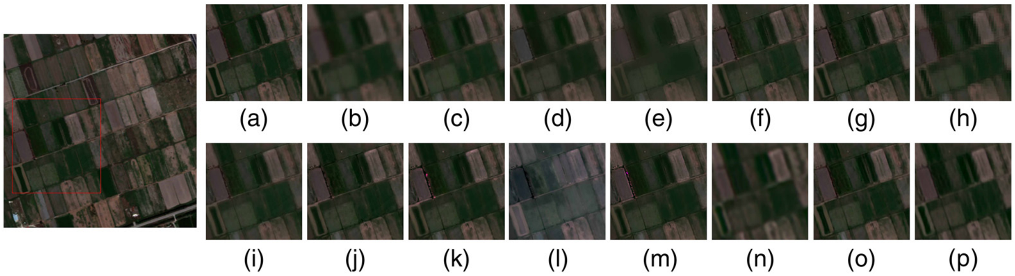

Fig. 13 Qualitative evaluation results on the GeoEye-1 dataset: (a) ground truth, (b) Bicubic, (c) Brovey, (d) CNMF, (e) GFPCA, (f) GS, (g) GSA, (h) wavelet, (i) HIS, (j) MTF_GLP, (k) MTF_GLP_HMP, (l) PCA, (m) SFIM, (n) PNN, (o) PanNet, and (p) ResTFNet.

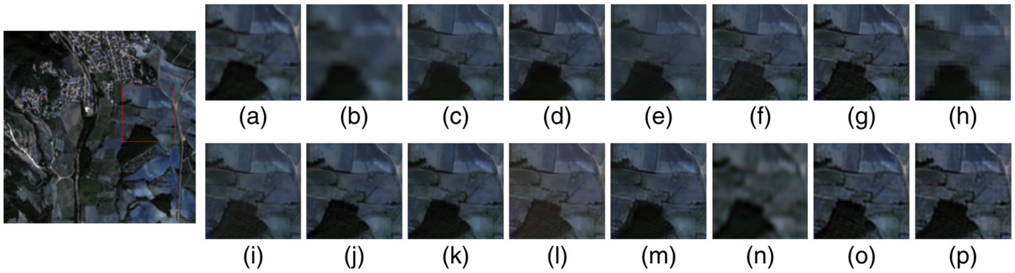

Fig. 14 Qualitative evaluation results on the Spot-6 dataset: (a) ground truth, (b) Bicubic, (c) Brovey, (d) CNMF, (e) GFPCA, (f) GS, (g) GSA, (h) wavelet, (i) HIS, (j) MTF_GLP, (k) MTF_GLP_HMP, (l) PCA, (m) SFIM, (n) PNN, (o) PanNet, and (p) ResTFNet.

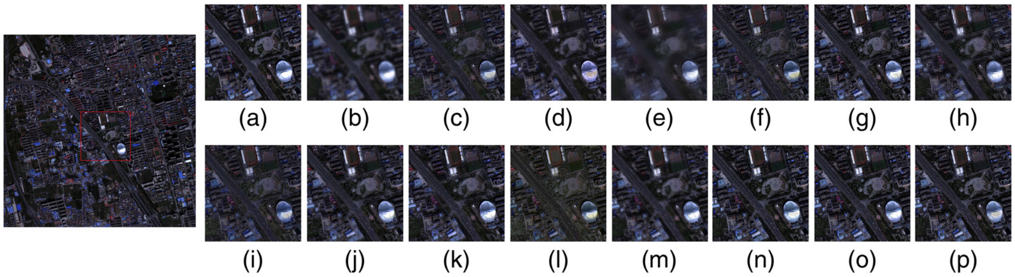

Fig. 15 Qualitative evaluation results on Gaofen-2 dataset: (a) ground truth, (b) Bicubic, (c) Brovey, (d) CNMF, (e) GFPCA, (f) GS, (g) GSA, (h) wavelet, (i) HIS, (j) MTF_GLP, (k) MTF_GLP_HMP, (l) PCA, (m) SFIM, (n) PNN, (o) PanNet, and (p) ResTFNet.

4.3.4 *Summary of comparative experiments*

In terms of reference evaluation measures, PanNet provides the best results for the three datasets because it combines knowledge of a specific field. To reconstruct spatial information, it trains the network parameters with high-frequency information. To reduce the spectrum usage, it directly propagates the upsampled multi-spectral image to the output of the network. Then high-pass filtering instead of spatial information reconstruction in the image domain provides proper generalization for images from different satellites. For the two-branch fusion ResTFNet, good results are achieved in terms of reference and no-reference evaluation measures. Hence, two-branch deep learning can outperform the direct stacking of inputs for reconstruction in space and spectrum of multi-frame remote sensing images. Deep learning is a new direction in the development of multi-frame remote sensing image super-division reconstruction algorithms in recent years, and it is indispensable in multi-frame remote sensing image super-division reconstruction algorithms due to its ability to characterize the non-linear relationship between observation data and its strong learning ability. Using deep learning can be regarded as the development direction of the field of multi-frame remote sensing image super-division reconstruction algorithm in the future, although the existing deep learning-based multi-frame remote sensing image super-division reconstruction method still has certain shortcomings. For example, in practical applications, there is still room for improvement in non-reference indicators. At the same time, similar to traditional methods, MS images are often regarded as the carrier of spectral information, while ignoring the spatial information of the images; and PAN images are regarded as the carrier of spatial information while ignoring their spectral information. These all lead to different degrees of information loss in the generated image in space and spectrum.

5 Summary and Outlook

Machine learning is becoming essential for remote sensing observations and analysis. Since the introduction of deep learning to process remote sensing images in the late 1990s, major achievements have been made in automatic feature extraction, land cover estimation, and oil spill monitoring. However, after a potential and comprehensive analysis of key remote sensing applications, various problems remain to be solved. For instance, the increase in available datasets and diversity of sampling methods complement spatiotemporal features of remote sensing images, but publicly available datasets are limited and have different sampling resolutions, resulting in a lack of high-quality remote sensing images, training samples, and ground truths. The following three points summarize the future directions and challenges of remote sensing image SR:

(1) Regarding spatiotemporal feature fusion of remote sensing images, the increasing number of datasets reduces the effectiveness of single-frame reconstruction to exploit complementarity between datasets. Therefore, multi-frame reconstruction has gradually become the research mainstream. Fully using datasets such as drone shooting data or topographic maps, land use classification maps, and sampling point data can contribute to the extraction and fusion of spatiotemporal features across datasets, thereby improving the reconstruction quality.

(2) Deep learning methods applied to remote sensing images should tend to lightweight frameworks. Algorithms based on deep learning can use deep networks to extract rich image information. Although deep learning can achieve a high performance, the calculation burden increases with the network complexity, and the hardware requirements become prohibitive for large-scale applications. Thus optimizing deep learning algorithms, adopting lightweight neural networks, and improving the computing efficiency can increase the applicability of SR reconstruction.

(3) Unsupervised learning is often used in SR reconstruction of remote sensing images because supervised learning usually requires labeled data to guide training, and few labeled datasets are available for SR reconstruction of remote sensing images. Considering fewer data labels may lead to convenient supervised learning for SR reconstruction of such images.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (Grant No. 61966035), by the International Cooperation Project of the Science and Technology Department of the Autonomous Region "Data-Driven Construction of Sino-Russian Cloud Computing Sharing Platform" (Grant No. 2020E01023), by the National Science Foundation of China (Grant No. U1803261), and by the Autonomous Region Graduate Innovation Project (No. XJ2021G062 and XJ2021G080).

References

1. Y. Zhang et al., "Remote sensing images super-resolution based on sparse dictionaries and residual dictionaries," in *IEEE 11th Int. Conf. Dependable, Autonomic and Secure Comput.*, IEEE, pp. 318–323 (2013).
2. W. Wu et al., "A new framework for remote sensing image super-resolution: sparse representation-based method by processing dictionaries with multi-type features," *J. Syst. Architect.* **64**, 63–75 (2016).
3. J. Amorós-López et al., "Multitemporal fusion of Landsat/TM and ENVISAT/MERIS for crop monitoring," *Int. J. Appl. Earth Obs. Geoinf.* **23**, 132–141 (2013).
4. J. Walker et al., "Evaluation of Landsat and MODIS data fusion products for analysis of dryland forest phenology," *Remote Sens. Environ.* **117**, 381–393 (2012).
5. M. A. White and R. R. Nemani, "Real-time monitoring and short-term forecasting of land surface phenology," *Remote Sens. Environ.* **104**(1), 43–49 (2006).
6. S. Gou et al., "Remote sensing image super-resolution reconstruction based on nonlocal pairwise dictionaries and double regularization," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **7**(12), 4784–4792 (2014).
7. D. Yang et al., "Remote sensing image super-resolution: challenges and approaches," in *IEEE Int. Conf. Digital Signal Process. (DSP)*, IEEE, pp. 196–200 (2015).
8. X. Qian et al., "Evaluation of the effect of feature extraction strategy on the performance of high-resolution remote sensing image scene classification," *J. Remote Sens.* **22**(5), 758–776 (2018).
9. J. Yang et al., "Image super-resolution via sparse representation," *IEEE Trans. Image Process.* **19**(11), 2861–2873 (2010).
10. S. Mei et al., "Spatial and spectral joint super-resolution using convolutional neural network," *IEEE Trans. Geosci. Remote Sens.* **58**(7), 4590–4603 (2020).
11. X. Dou et al., "Super-resolution for hyperspectral remote sensing images based on the 3D attention-SRGAN network," *Remote Sens.* **12**(7), 1204 (2020).
12. K. Zhang et al., "Learning multiple linear mappings for efficient single image super-resolution," *IEEE Trans. Image Process.* **24**(3), 846–861 (2015).
13. L. Li et al., "Super-resolution reconstruction of high-resolution satellite ZY-3 TLC images," *Sensors* **17**(5), 1062 (2017).
14. J. L. Harris, "Diffraction and resolving power," *J. Opt. Soc. Am.* **54**(7), 931–936 (1964).
15. E. Huggins, "Introduction to Fourier optics," *Phys. Teach.* **45**(6), 364–368 (2007).

16. R. Tsai, "Multiframe image restoration and registration," *Adv. Comput. Vis. Image Process.* **1**, 317–339 (1984).
17. H. Shen et al., "Super-resolution reconstruction algorithm to MODIS remote sensing images," *Comput. J.* **52**(1), 90–100 (2009).
18. R. Wang et al., "Lightweight non-local network for image super-resolution," in *ICASSP 2021-2021 IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP)*, IEEE, pp. 1625–1629 (2021).
19. H. Tao et al., "Superresolution remote sensing image processing algorithm based on wavelet transform and interpolation," *Proc. SPIE* **4898**, 259–263 (2003).
20. Z. H. Wei et al., "A wavelet-based restoration algorithm of remote sensing images," in *6th Natl. Joint Acad. Conf. Signal and Inf. Process*, pp. 182–184 (2007).
21. W. Jinliang et al., "Super resolution reconstruction of tm remote sensing image based on wavelet analysis," *Remote Sens. Technol. Appl.* **31**(3), 476–480 (2016).
22. W. Ma et al., "Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive Res-Net," *IEEE Trans. Geosci. Remote Sens.* **57**(6), 3512–3527 (2019).
23. W. Jingmeng et al., "Super-resolution reconstruction of remote sensing image based on staggered pixels and non-uniform B-spline curved surface," *Remote Sens. Land Resour.* **27**(1), 35–43 (2015).
24. S. Y. Han, N. H. Park, and K. H. Joo, "Wavelet transform based image interpolation for remote sensing image," *Int. J. Software Eng. Appl.* **9**(2), 59–66 (2015).
25. P. Solanki, D. Israni, and A. Shah, "An efficient satellite image super resolution technique for shift-variant images using improved new edge directed interpolation," *Stat. Optim. Inf. Comput.* **6**(4), 619–632 (2018).
26. L. Shang, S.-F. Liu, and Z.-L. Sun, "Image super-resolution reconstruction based on sparse representation and POCS method," in *Int. Conf. Intell. Comput.*, Springer, pp. 348–356 (2015).
27. A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time," *IEEE Trans. Image Process.* **6**(8), 1064–1076 (1997).
28. A. J. Patti, M. I. Sezan, and A. M. Tekalp, "Robust methods for high-quality stills from interlaced video in the presence of dominant motion," *IEEE Trans. Circuits Syst. Video Technol.* **7**(2), 328–342 (1997).
29. L. Tao, Q. Feng, and Z. Bao, "Map super-resolution reconstruction of remote sensing image," *Chin. J. Liquid Cryst. Disp.* **33**(10), 884–892 (2018).
30. H. Irmak et al., "Super-resolution reconstruction of hyperspectral images via an improved map-based approach," in *IEEE Int. Geosci. and Remote Sens. Symp. (IGARSS)*, IEEE, pp. 7244–7247 (2016).
31. H. Irmak, G. B. Akar, and S. E. Yuksel, "A map-based approach for hyperspectral imagery super-resolution," *IEEE Trans. Image Process.* **27**(6), 2942–2951 (2018).
32. T. Guo and W. Song, "Super resolution reconstruction of remote sensing image based on high frequency enhancement curve and iterative back projection," *Eng. Surv. Mapp.* **1**, 64–67, 72 (2018).
33. F. R. Deepa and M. J. Islam, "Effect of atmospheric turbulence on the performance of underwater wireless SAC-OCDMA system," in *2nd Int. Conf. Sustain. Technol. for Ind. 4.0 (STI)*, IEEE, pp. 1–5 (2020).
34. X. Li et al., "Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning," *IEEE Trans. Geosci. Remote Sens.* **52**(11), 7086–7098 (2014).
35. F. Li, D. Fraser, and X. Jia, "Improved IBP for super-resolving remote sensing images," *Geogr. Inf. Sci.* **12**(2), 106–111 (2006).
36. G. Tongyu, "Super resolution reconstruction of remote sensing image based on improved iterative back projection algorithm," *Geomat. Spatial Inf. Technol.* **42**(1), 195–197, 205 (2019).
37. L. Lu et al., "A method of images super-resolution reconstruction based on MRF-map frame," *J. Xiamen Univ. (Nat. Sci.)* **4** (2012).

38. W. T. Freeman, T. R. Jones, and E. C. Pasztor, "Example-based super-resolution," *IEEE Comput. Graph. Appl.* **22**(2), 56–65 (2002).
39. F.-L. Wu and X.-J. Wang, "Example-based super-resolution for single-image analysis from the chang'e-1 mission," *Res. Astron. Astrophys.* **16**(11), 172 (2016).
40. K. Su et al., "Neighborhood issue in single-frame image super-resolution," in *IEEE Int. Conf. Multimedia and Expo*, IEEE, pp. 1–4 (2005).
41. H. Chang, D.-Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," in *Proc. 2004 IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit., 2004. CVPR 2004*, IEEE, Vol. 1, pp. I–I (2004).
42. W. Xinlei and L. Naifeng, "Super-resolution of remote sensing images via sparse structural manifold embedding," *Neurocomputing* **173**, 1402–1411 (2016).
43. T. Zhang, Y. Du, and F. Lu, "Super-resolution reconstruction of remote sensing images using multiple-point statistics and isometric mapping," *Remote Sens.* **9**(7), 724 (2017).
44. Z. Zhihui, W. Bo, and S. Kang, "Single remote sensing image super-resolution and denoising via sparse representation," in *Int. Workshop Multi-Platform/Multi-Sens. Remote Sens. and Mapp.*, IEEE, pp. 1–5 (2011).
45. S. Liu, Y. Zhu, and L. Xue, "Remote sensing image super-resolution method using sparse representation and classified texture patches," *Wuhan Daxue Xuebao* **40**(5), 578–582 (2015).
46. K. Zhang et al., "Joint learning of multiple regressors for single image super-resolution," *IEEE Signal Process Lett.* **23**(1), 102–106 (2015).
47. Y. Zhang et al., "Improvement of the example-regression-based super-resolution land cover mapping algorithm," *IEEE Geosci. Remote Sens. Lett.* **12**(8), 1740–1744 (2015).
48. W. Jin et al., "The improvements of BP neural network learning algorithm," in *WCC 2000-ICSP 2000. 2000 5th Int. Conf. Signal Process. Proc. 16th World Comput. Congr. 2000*, IEEE, Vol. 3, pp. 1647–1649 (2000).
49. H. Y. Ding, "Remote sensed image super-resolution reconstruction based on BP neural network," *Comput. Eng. Appl.* **44**(1), 171–84 (2008).
50. W. Chen and Y.-F. Wang, "A study on a new method of multi-spatial-resolution remote sensing image fusion based on GA–BP," *Remote Sens. Technol. Appl.* **22**(4) (2007).
51. D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *J. Physiol.* **160**(1), 106–154 (1962).
52. Y. LeCun et al., "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).
53. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Adv. Neural Inf. Process. Syst.* **25**, 1097–1105 (2012).
54. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *ICML* (2010).
55. A. Ducournau and R. Fablet, "Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived SST data," in *9th IAPR Workshop Pattern Recognit. in Remote Sens. (PRRS)*, IEEE, pp. 1–6 (2016).
56. C. Dong et al., "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(2), 295–307 (2015).
57. S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local–global combined network," *IEEE Geosci. Remote Sens. Lett.* **14**(8), 1243–1247 (2017).
58. G. Masi et al., "Pansharpening by convolutional neural networks," *Remote Sens.* **8**(7), 594 (2016).
59. F. Ye, X. Li, and X. Zhang, "FusionCNN: a remote sensing image fusion algorithm based on deep convolutional neural networks," *Multimedia Tools Appl.* **78**(11), 14683–14703 (2019).
60. F. Palsson, J. R. Sveinsson, and M. O. Ulfarsson, "Multispectral and hyperspectral image fusion using a 3-D-convolutional neural network," *IEEE Geosci. Remote Sens. Lett.* **14**(5), 639–643 (2017).
61. J. Yang, Y.-Q. Zhao, and J. C.-W. Chan, "Hyperspectral and multispectral image fusion via deep two-branches convolutional neural network," *Remote Sens.* **10**(5), 800 (2018).
62. I. J. Goodfellow et al., "Generative adversarial networks," *Commun. ACM* **63**(11), 139–144 (2020).

63. M. Arjovsky and L. Bottou, “Towards principled methods for training generative adversarial networks,” https://arxiv.org/abs/1701.04862 (2017).
64. A. Radford, L. Metz, and S. Chintala, “Unsupervised representation learning with deep convolutional generative adversarial networks,” https://arxiv.org/abs/1511.06434 (2015).
65. X. Chen et al., “InfoGAN: interpretable representation learning by information maximizing generative adversarial nets,” 2180–2188, https://arxiv.org/abs/1606.03657 (2016).
66. J. Zhao, M. Mathieu, and Y. LeCun, “Energy-based generative adversarial network,” https://arxiv.org/abs/1609.03126 (2016).
67. X. Mao et al., “Least squares generative adversarial networks,” in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2794–2802 (2017).
68. C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4681–4690 (2017).
69. Q. Huang et al., “Hyperspectral image super-resolution using generative adversarial network and residual learning,” in *ICASSP 2019-2019 IEEE Int. Conf. Acous. Speech and Signal Process. (ICASSP)*, IEEE, pp. 3012–3016 (2019).
70. Q. Liu et al., “PSGAN: a generative adversarial network for remote sensing image PAN-sharpening,” *IEEE Trans. Geosci. Remote Sens.* 1–16(2020).
71. Z. Shao et al., “Residual encoder–decoder conditional generative adversarial network for pansharpening,” *IEEE Geosci. Remote Sens. Lett.* **17**(9), 1573–1577 (2019).
72. J. Ma et al., “PAN-GAN: an unsupervised PAN-sharpening method for remote sensing image fusion,” *Inf. Fusion* **62**, 110–120 (2020).
73. G. Huang et al., “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 4700–4708 (2017).
74. Z. Pan et al., “Super-resolution of single remote sensing image based on residual dense backprojection networks,” *IEEE Trans. Geosci. Remote Sens.* **57**(10), 7918–7933 (2019).
75. K. He et al., “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 770–778 (2016).
76. B. Lim et al., “Enhanced deep residual networks for single image super-resolution,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, pp. 136–144 (2017).
77. A. Wang, X. Song, and Y. Chen, “Super-resolution reconstruction of remote sensing image based on recursive residual network,” *Comput. Eng. Appl.* **55**(3), 191–195 (2019).
78. Y. Yang et al., “PCDRN: progressive cascade deep residual network for pansharpening,” *Remote Sens.* **12**(4), 676 (2020).
79. Y. Zheng et al., “Deep residual learning for boosting the accuracy of hyperspectral pansharpening,” *IEEE Geosci. Remote Sens. Lett.* **17**(8), 1435–1439 (2019).
80. Y. Zheng et al., “Hyperspectral pansharpening using deep prior and dual attention residual network,” *IEEE Trans. Geosci. Remote Sens.* **58**(11), 8059–8076 (2020).
81. J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 7132–7141 (2018).
82. Y. Zhang et al., “Image super-resolution using very deep residual channel attention networks,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 286–301 (2018).
83. J. M. Haut et al., “Remote sensing image superresolution using deep residual channel attention,” *IEEE Trans. Geosci. Remote Sens.* **57**(11), 9277–9289 (2019).
84. S. Woo et al., “CBAM: convolutional block attention module,” in *Proc. Eur. Conf. Comput. Vision (ECCV)*, pp. 3–19 (2018).
85. J. Liu et al., “Residual feature aggregation network for image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 2359–2368 (2020).
86. Q. Yang et al., “Hyperspectral and multispectral image fusion based on deep attention network,” in *10th Workshop Hyperspectral Imaging and Signal Process.: Evol. in Remote Sens. (WHISPERS)*, IEEE, pp. 1–5 (2019).
87. Q. Huynh-Thu and M. Ghanbari, “Scope of validity of PSNR in image/video quality assessment,” *Electron. Lett.* **44**(13), 800–801 (2008).
88. Z. Chen et al., “Fusion of hyperspectral and multispectral images: a novel framework based on generalization of PAN-sharpening methods,” *IEEE Geosci. Remote Sens. Lett.* **11**(8), 1418–1422 (2014).

89. L. Wald, "Quality of high resolution synthesised images: is there a simple criterion?" in *Third Conf. "Fusion of Earth Data: Merging Point Meas., Raster Maps and Remotely Sens. Images"*, SEE/URISCA, pp. 99–103 (2000).
90. R. H. Yuhas, A. F. Goetz, and J. W. Boardman, "Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm," in *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop*, Vol. **1**, pp. 147–149 (1992).
91. L. Alparone et al., "Comparison of pansharpening algorithms: outcome of the 2006 GRS-S data-fusion contest," *IEEE Trans. Geosci. Remote Sens.* **45**(10), 3012–3021 (2007).
92. L. Wald, T. Ranchin, and M. Mangolini, "Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images," *Photogramm. Eng. Remote Sens.* **63**(6), 691–699 (1997).
93. L. Alparone et al., "Multispectral and panchromatic data fusion assessment without reference," *Photogramm. Eng. Remote Sens.* **74**(2), 193–200 (2008).
94. C. Li et al., "No-training, no-reference image quality index using perceptual features," *Opt. Eng.* **52**(5), 057003 (2013).
95. Y. Zhang et al., "Residual dense network for image super-resolution," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 2472–2481 (2018).
96. Y. Guo et al., "Closed-loop matters: dual regression networks for single image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, pp. 5407–5416 (2020).
97. N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: further improving enhanced super-resolution generative adversarial network," in *ICASSP 2020-2020 IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP)* (2020).
98. V. K. Shettigara, "A generalized component substitution technique for spatial enhancement of multispectral images using a higher resolution data set," *Photogramm. Eng. Remote Sens.* **58**(5), 561–567 (1992).
99. W. Carper, T. Lillesand, and R. Kiefer, "The use of intensity-hue-saturation transformations for merging spot panchromatic and multispectral image data," *Photogramm. Eng. Remote Sens.* **56**(4), 459–467 (1990).
100. A. R. Gillespie, A. B. Kahle, and R. E. Walker, "Color enhancement of highly correlated images. II. Channel ratio and "chromaticity" transformation techniques," *Remote Sens. Environ.* **22**(3), 343–365 (1987).
101. J. Liu, "Smoothing filter-based intensity modulation: a spectral preserve image fusion technique for improving spatial details," *Int. J. Remote Sens.* **21**(18), 3461–3472 (2000).
102. C. A. Laben and B. V. Brower, "Process for enhancing the spatial resolution of multispectral imagery using PAN-sharpening," U.S. Patent (Harris Corp.) No. 6,011,875 (2000).
103. R. L. King and J. Wang, "A wavelet based algorithm for PAN sharpening Landsat 7 imagery," in *IGARSS 2001. Scanning the Present and Resolving the Future. Proc. IEEE 2001 Int. Geosci. and Remote Sens. Symp. (Cat. No. 01CH37217)*, IEEE, Vol. 2, pp. 849–851 (2001).
104. B. Aiazzi et al., "MTF-tailored multiscale fusion of high-resolution MS and PAN imagery," *Photogramm. Eng. Remote Sens.* **72**(5), 591–596 (2006).
105. B. Aiazzi et al., "An MTF-based spectral distortion minimizing model for PAN-sharpening of very high resolution multispectral images of urban areas," in *2nd GRSS/ISPRS Joint Workshop Remote Sens. and Data Fusion over Urban Areas*, IEEE, pp. 90–94 (2003).
106. B. Aiazzi, S. Baronti, and M. Selva, "Improving component substitution pansharpening through multivariate regression of MS + PAN data," *IEEE Trans. Geosci. Remote Sens.* **45**(10), 3230–3239 (2007).
107. N. Yokoya, T. Yairi, and A. Iwasaki, "Coupled nonnegative matrix factorization unmixing for hyperspectral and multispectral data fusion," *IEEE Trans. Geosci. Remote Sens.* **50**(2), 528–537 (2011).
108. W. Liao et al., "Two-stage fusion of thermal hyperspectral and visible RGB image by PCA and guided filter," in *7th Workshop Hyperspectral Image and Signal Process.: Evol. in Remote Sens. (WHISPERS)*, IEEE, pp. 1–4 (2015).
109. J. Yang et al., "PanNet: a deep network architecture for PAN-sharpening," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 5449–5457 (2017).

110. X. Liu, Q. Liu, and Y. Wang, “Remote sensing image fusion based on two-stream fusion network,” *Inf. Fusion* **55**, 1–15 (2020).

Hui Liu received her BS and MS degrees in software engineering from Xinjiang University in 2014 and 2017. She is currently pursuing her PhD in computer science and technology, Xinjiang University, Urumqi, China. She is a PhD candidate at Xinjiang University. Her research interests include deep learning and opportunistic networks and processing of remote sensing image data.

Yurong Qian received her bachelor’s and master’s degrees in computer science and technology from Xinjiang University in 2000 and her doctorate degree in biology from Nanjing University in 2010. She is currently a professor at the School of Software, Xinjiang University, Urumqi, China. Her research interests include computational intelligence such as big data processing, image processing, and artificial neural networks.

Xiwu Zhong received his bachelor’s degree in network engineering from Huizhou University in 2018. Currently, he is working toward his master’s degree majoring in software engineering at the Xinjiang University. His research interests include deep learning and remote sensing image processing.

Long Chen graduated from Shandong University of Science and Technology with his bachelor’s degree in geographic information science in 2018. At Xinjiang University, he is currently pursuing his master’s degree in software engineering. Deep learning and single image super resolution are two of his research interests.

Guangqi Yang obtained his bachelor’s degree in software engineering from Changshu Institute of Technology in 2020 and is currently studying for his master’s degree in software engineering at Xinjiang University, Urumqi, China. His research interests include deep learning and remote sensing image data processing.