Super-resolution method using generative adversarial network for Gaofen wide-field-view images

Ziyun Zhang,[a] Chengming Zhang,[a,*] Menxin Wu,[b] Yingjuan Han,[c] Hao Yin,[a] Ailing Kong,[a] and Fangfang Chen[a]

[a]Shandong Agricultural University, College of Information Science and Engineering, Taian, China

[b]National Meteorological Center/CMA, Beijing, China

[c]Key Laboratory for Meteorological Disaster Monitoring and Early Warning and Risk Management of Characteristic Agriculture in Arid Regions, CMA, Yinchuan, China

Abstract. Accurate information on the spatial distribution of crops is of great significance for scientific research and production practices. Such accurate information can be extracted from high-spatial-resolution optical remote sensing images. However, acquiring these images with a wide coverage is difficult. We established a model named multispectral super-resolution generative adversarial network (MS_SRGAN) for generating high-resolution 4-m images using Gaofen 1 wide-field-view (WFV) 16-m images. The MS_SRGAN model contains a generator and a discriminator. The generator network is composed of feature extraction units and feature fusion units with a symmetric structure, and the attention mechanism is introduced to constrain the spectral value of the feature map during feature extraction. The generator loss introduces feature loss to describe the feature difference of the image. This is realized using pre-trained discriminator parameters and a partial discriminator network. In addition to realizing feature loss, the discriminator network, which is a simple convolutional neural network, also realizes adversarial loss. Adversarial loss can provide some fake high frequency details to the generator to get a more sharpened image. In the Gaofen 1 WFV image test, the performance of MS_SRGAN was compared with that of Bicubic, EDSR, SRGAN, and ESRGAN. The results show that the spectral angle mapper (3.387) and structural similarity index measure (0.998) of MS_SRGAN are higher than those of the other models. In addition, the image obtained by MS_SRGAN is more realistic; its texture details and color distribution are closer to the reference image to a greater extent. [DOI: 10.1117/1.JRS.15.028506]

Keywords: super-resolution; generative adversarial network; multispectral image; Gaofen 1 WFV image; Gaofen 2 image; convolutional neural network.

Paper 200877 received Dec. 15, 2020; accepted for publication Jun. 9, 2021; published online Jun. 25, 2021.

1 Introduction

Remote sensing technology enables the acquisition of ground information over large areas. Accordingly, remote sensing images have become an important source of basic data in many fields such as global change studies, agricultural monitoring, and resource and environmental surveys. Satellite imagery can be used for crop information extraction over large-scale areas. With the advantages of wide spatial coverage, rich spectral information, and high temporal resolution, wide-field-view (WFV) images from Chinese satellites Gaofen-1 (GF1) and Gaofen-6 (GF6) have a high application potential for crop information extraction. However, they cannot be directly used to extract detailed information on the spatial distribution of crops due to their lower spatial resolution of 16 m. In contrast, panchromatic and multispectral sensor (PMS) images of Gaofen-2 (GF2) have a spatial resolution of 1 and 4 m, respectively, offering the opportunity to extract accurate information on the spatial distribution of crops. However, their application is

*Address all correspondence to Chengming Zhang, chming@sdau.edu.cn

limited by a narrow spatial coverage and low temporal resolution, which are not favorable for large-scale crops.[1] Therefore, super-resolution—an image reconstruction technique[2]—of GF1 and GF6 images with GF2 images as reference data is of high significance for extracting accurate information on the spatial distribution of crops over a wide area.

Super-resolution can be mainly classified into two types: single image super-resolution (SISR)[3] and multi-frame super-resolution[4] Classic SISR methods include interpolation,[5] maximum a posteriori probability (MAP),[6,7] and projections onto convex sets of algorithms.[8] Most of these classical methods are based on statistical analysis. Recently, researchers have introduced machine learning to super-resolution[9] such that improved algorithms can acquire more information to improve the quality of the generated images. Super-resolution algorithms that are based on dictionary learning,[10–12] local linear regression,[13,14] and neural networks have shown positive results.

Convolutional neural networks (CNNs) have strong autonomous learning capabilities and outstanding advantages in feature extraction.[15–19] By fully integrating the advantages of CNNs in feature extraction, super-resolution CNNs (SRCNN) can generate super-resolution images by adjusting high-resolution images reconstructed using the Bicubic interpolation method.[20] When applying the advantages of CNNs, researchers have built various super-resolution networks, including very deep convolutional networks (VDSR),[21] residual encoder-decoder networks,[22] deeply recursive convolutional networks,[23] Laplacian pyramid super-resolution networks,[24] super-resolution DenseNet (SRDenseNet),[25] enhanced deep residual networks (EDSR),[26] and residual channel attention networks (RCAN).[27] Among them, VDSR, SRDenseNet, and RCAN use the feature extraction method of image classification to deepen the network; the effectiveness of this network structure in super-resolution has been proven through experiments. In terms of upsampling, methods based on convolution (deconvolution[28] and pixel-shuffle pixel[29] methods) clearly show a higher performance than the methods based on interpolation.

Generative adversarial networks (GANs)[30] have shown excellent results in various fields, such as image style migration,[31–33] super-resolution image completion,[34–36] and denoising.[37–39] GANs have some advantages in super-resolution because a discriminator network, which is introduced in them, uses two networks to train each other and enables the discriminator to instruct the generator to produce an image with enhanced high-frequency textural detail. Super-resolution GANs (SRGAN),[40] which are based on the retention of traditional loss, significantly improve the effect of image generation by further adding perceptual loss.[41] Perceptual loss uses a pre-trained network to extract a feature map that can reflect the overall structure of images and calculate the Euclidean distance between low-resolution and high-resolution feature maps. This multi-loss joint mechanism strengthens the optimization ability of the generated network and enables the reconstruction of the overall structural features of high-resolution images. According to the literature, SRGAN exhibits a distinctly stronger sharpening effect than SRResNet and other methods. The enhanced super-resolution GAN (ESRGAN) combines the residual structure and dense connectivity to introduce a residual-in-residual dense block (RRDB), which enhances the feature mapping capability of the generator and further improves the accuracy of the results.[42] RankSRGAN adds a loss of rank content based on the standard SRGAN loss, which enhances the training ability of the generator. Previous studies have reported that the loss of rank content can be applied to various methods and can further improve the accuracy of the results.[43,44]

At present, the technology of super-resolution is mature for natural images. Unlike natural images, remote sensing images have more channels and less pixel details. Therefore, a new super-resolution technology based on the features of remote sensing images is required. A previous study used deep-connectivity and residual-connections SRCNN (DCR_SRCNN) with a Sentinel-2 image as a reference to realize super-resolution of Landsat images. The experimental results showed that super-resolution was strongly affected by an excessively long time interval between the low-resolution images and the reference images in the dataset.[45] The extended super-resolution convolutional neural network uses Landsat-8 and Sentinel-2 images at different moments and overcomes the limitation of temporal resolution to achieve multitemporal image fusion.[46] The progressive residual depth neural network makes super resolution of the DOTA satellite image database. Here, the progressive residual structure is used to find the feature

information of remote sensing images at different levels to provide more detailed features for the reconstruction of super-resolution remote sensing images.[47] The dense residual generative adversarial network organically combines a dense connection structure with a residual structure to form a generating network. The Wasserstein GAN-gradient penalty (WGAN-GP) adversarial loss calculation method has been adopted in this paper. Many experiments on the NWPU-RESISC45 dataset show that this method can further improve the accuracy of the model in the super-resolution of remote sensing images.[48]

We aimed to achieve super-resolution of GF1 and GF6 WFV images to generate images of higher spatial resolution. A GF2 PMS image was used as the reference image, and a GAN model was used to establish a method for the super-resolution of the WFV images. This method is called multispectral super-resolution GAN (MS_SRGAN). The main contributions of this study are as follows:

1. We introduced a residual squeeze-excitation (RSE) block to adjust the data distribution in the generated image to solve the problem of inconsistency between the distribution of Gaofen WFV data and reference image data. Furthermore, we established a generation network with the RSE block that extracts the features of different levels of the image and fuses the corresponding low-level features with high-level features to further improve the accuracy of the generated image.
2. We added feature loss to describe the difference in the image features, which is realized by the partial discriminator network, to account for generator loss.

2 Study Area and Dataset

2.1 *Study Area*

We selected the Shandong Province and the Ningxia Hui Autonomous Region as the study areas. Shandong is a major agricultural province in China, with wheat, corn, and sweet potato as the major crops. It covers an area of 157,900 km^2 (34°22′–38°24′N, 114°47′–122°42′E), and its grain output accounts for 8.1% of the national output. The Ningxia Hui Autonomous Region covers an area of 66,400 km^2 (35°14′–39°23′N, 104°17′–107°39′E). It has the agricultural characteristics of northwest China, and the main grain crops of this region are maize and wheat.

We collected images covering the flat areas of northwest and southwest Shandong and the south-central plain of Ningxia. As the investigation was focused on cropland, most of the selected images mainly feature cropland (Fig. 1).

2.2 *Dataset*

In this study, 18 GF2 PMS images from June 2019, 16 GF2 PMS images from March 2020, eight GF1 WFV images from June 2019, and five GF6 WFV images from March 2020 were collected. Due to the large amount of data used in this study, only part of the image information is shown in Table 1 as a representative sample. Each GF2 and GF1 image contains multispectral bands [red, green, blue, and near-infrared (NIR)], and the GF6 images also contain multispectral bands (red, green, blue, NIR, red edge, red edge 2, water body blue, and yellow).

First, we identified low-resolution remote sensing images and then selected different types of high-resolution remote sensing images to act as references to build the dataset. During the selection, the following three aspects were considered: the coverage of low-resolution images, the size of the super-resolution factor (the ratio of low-resolution images to high-resolution images), and the band range.

Gaofen WFV images have a width of 800 km, which represents a high-quality medium-spatial resolution. After networking, the revisit cycles of the GF1 and GF6 satellites were shortened to two days, thereby providing satisfactory temporal resolution and coverage. Tables 2 and 3 shows the main parameters of the GF1 and GF6 images, respectively.

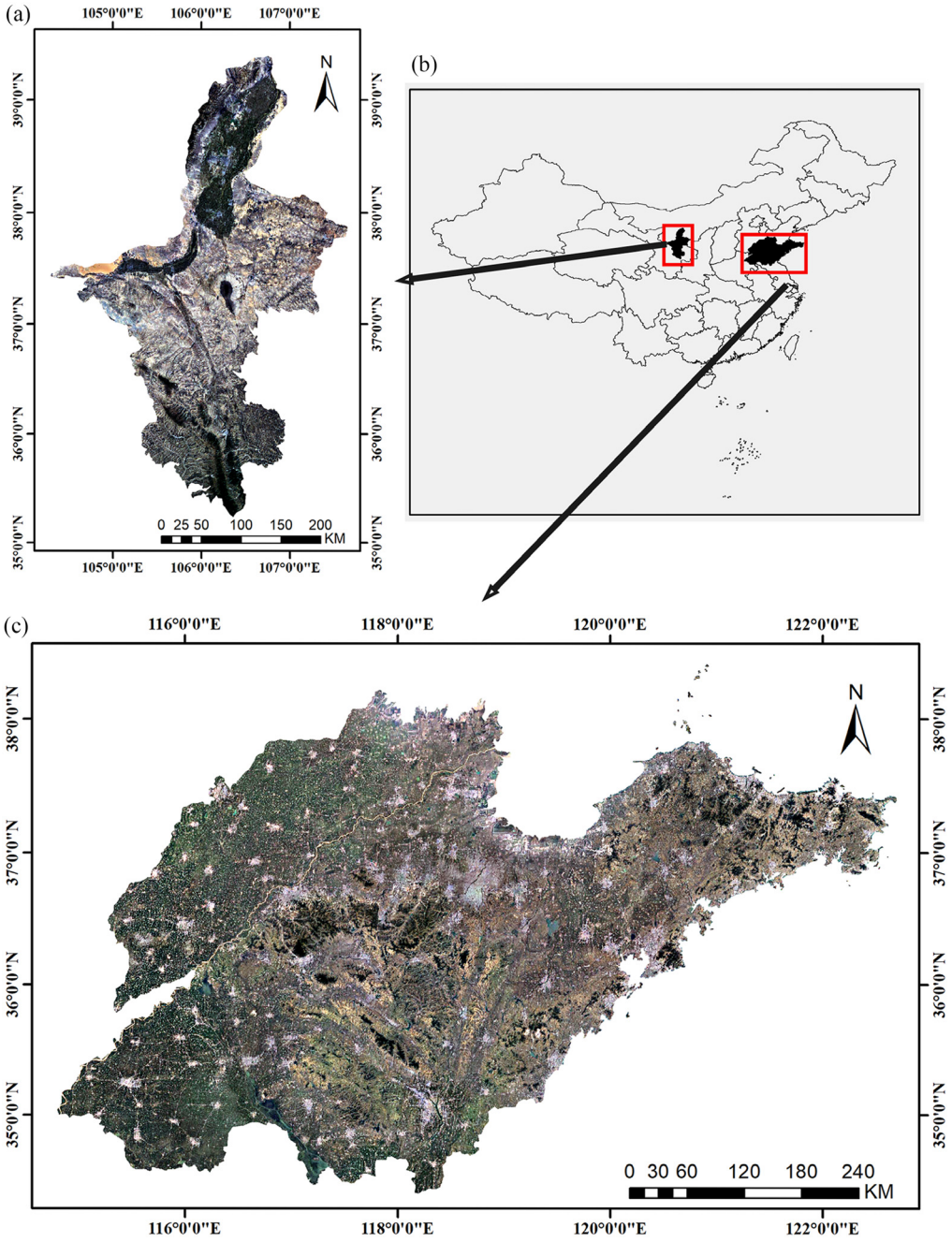

Fig. 1 (a) Map showing the location of the study areas, and (b) satellite images of the Ningxia Hui Autonomous Region and (c) Shandong Province.

The GF2 PMS images have a spatial resolution of 4 m, which is consistent with the band range of the Gaofen WFV images. At the same time, it can be seen from the main parameters of the GF2 image in Table 4 that the scale range and temporal resolution of the GF2 image are inferior to the Gaofen WFV image, and the super-resolution of the GF2 image and Gaofen WFV image can greatly supplement the image with 4m spatial resolution.

In addition, the multispectral images selected in this experiment include the red, green, blue, and NIR bands. The spectral range of the NIR band is included in the red band, which can highlight the textural features of crops and coincides with the goal of our super-resolution of crop images.

We used remote sensing image processing software to preprocess images such as atmospheric correction, radiometric correction, and geographical registration such that all of the

Table 1 Partial image information.

Location	Sensor	Latitude and longitude	Time	Filename
Ningxia	GF1-WFV3	E106.5,N35.6	20190603	GF1_WFV3_E106.5_N35.6_20190603_L1A0004039767
Ningxia	GF1-WFV3	E107.0,N37.3	20190603	GF1_WFV3_E107.0_N37.3_20190603_L1A0004039768
Ningxia	GF1-WFV3	E108.5,N38.9	20190603	GF1_WFV3_E108.5_N38.9_20190603_L1A0004039769
Ningxia	GF2-PMS1	E106.0,N36.0	20190605	GF2_PMS1_E106.0_N36.0_20190605_L1A0004043646
Ningxia	GF2-PMS1	E106.1,N36.3	20190605	GF2_PMS1_E106.1_N36.3_20190605_L1A0004043653
Ningxia	GF2-PMS1	E106.1,N36.5	20190605	GF2_PMS1_E106.1_N36.5_20190605_L1A0004043650
Shandong	GF6-WFV	E116.2,N37.2	20200304	GF6_WFV_E116.2_N37.2_20200304_L1A1119973192
Shandong	GF6-WFV	E116.4,N38.0	20200304	GF6_WFV_E116.4_N38.0_20200304_L1A1119973184
Shandong	GF6-WFV	E115.8,N36.0	20200304	GF6_WFV_E115.8_N36.0_20200304_L1A1119973194
Shandong	GF2-PMS2	E115.9,N36.5	20200310	GF2_PMS2_E115.9_N36.5_20200310_L1A0004666684
Shandong	GF2-PMS2	E116.0,N37.0	20200310	GF2_PMS2_E116.0_N37.0_20200310_L1A0004666703
Shandong	GF2-PMS2	E115.2,N35.1	20200310	GF2_PMS2_E115.2_N35.1_20200310_L1A0004664717

Table 2 Main parameters of GF1 image.

Parameters	16 m–MS sensor	
Spectral range	Multispectral	0.45 to 0.52 μm
		0.52 to 0.59 μm
		0.63 to 0.69 μm
		0.77 to 0.89 μm
Spatial resolution	Multispectral	16 m
Scale range	800 km	
Revisit cycle	4 days	

images are placed under the same geographic coordinate system. Then, the preprocessed reference image was cut into 480×480 pixels corresponding to the low-resolution image that was cut into 120×120 pixels. The reference and low-resolution images contain four channels of red, blue, green, and NIR. We stripped the bands of the GF6 WFV images that were not included in the GF2 images to ensure compatibility among the image bands.

Finally, we obtained 1300 pairs of image blocks in the GF1–GF2 dataset and 1600 pairs of image blocks in the GF6–GF2 dataset. Among them, 60% of the image block pairs in the two data sets were used for training, 10% for validation, and 30% for testing.

Table 3 Main parameters of GF6 image.

Parameters	16 m–MS sensor	
Spectral range	Multispectral	0.45 to 0.52 μm
		0.52 to 0.59 μm
		0.63 to 0.69 μm
		0.77 to 0.89 μm
		0.69 to 0.73 μm
		0.73 to 0.77 μm
		0.40 to 0.45 μm
		0.59 to 0.63 μm
Spatial resolution	Multispectral	16 m
Scale range	800 km	
Revisit cycle	4 days	

Table 4 Main parameters of GF2 image.

Parameters	2 m–PAN sensor/8 m–MS sensor (μm)	
Spectral range	Panchromatic	0.45 to 0.90
	Multispectral	0.45 to 0.52
		0.52 to 0.59
		0.63 to 0.69
		0.77 to 0.89
Spatial resolution	Panchromatic	1 m
	Multispectral	4 m
Scale range	45 km	
Revisit cycle	5 days	

3 Method

3.1 *Key Question*

The GF1-GF2 and GF6-GF2 datasets were produced in this study according to the method in Pouliot et al.[45] The production method of this data set makes use of a known high-resolution image to carry out super-resolution for another low-resolution image, making full use of the advantages of the abundant data sources of remote sensing images. The known high-resolution images provide rich high-frequency details to the algorithm. However, different satellites use different instruments with various sensor ranges, so images of the same location sensed with different satellites might deviate from each other. It can be seen from Fig. 2 that, although the ground cover features of images GF1 and GF2 are the same, the actual point distribution histogram is quite different (Fig. 3).

This leads to the problem of inconsistent spectral distributions between the low-resolution and reference images, which makes feature extraction more difficult. To reduce the impact of this

(a) (b)

Fig. 2 (a) GF1 image and (b) GF2 image.

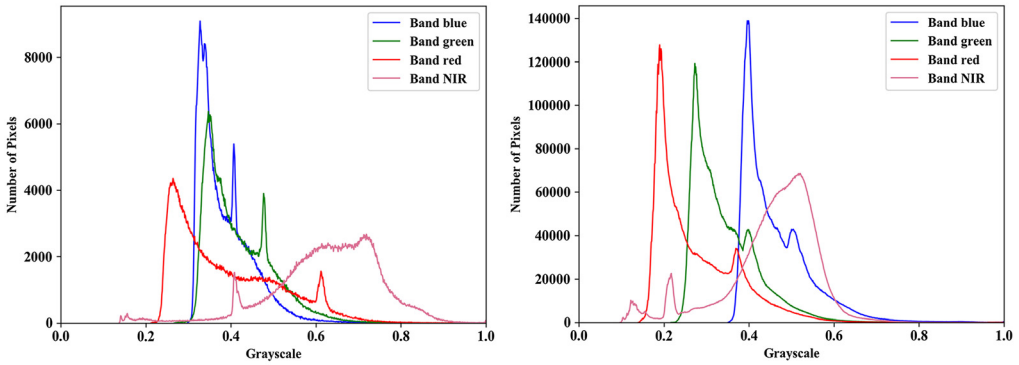

Fig. 3 Point distribution histogram of (a) GF1 image and (b) GF2 image.

problem, we chose to add RSE blocks to the generator to enhance its simulation ability and further improve the similarity of texture details between the generated image and the reference image.

3.2 *Structure of MS_SRGAN*

The structure of the proposed model is shown in Fig. 4. The generator network comprises an RSE block, a convolutional layer, and a deconvolutional layer. The generator loss constitutes three parts: adversarial loss, per-pixel loss, and feature loss. The discriminator network consists of a convolutional layer, global average pooling, and an activation layer. The discriminator loss is realized by the Wasserstein distance.

3.3 *Generator*

Considering the problem described in Sec. 3.1, we introduce an attention mechanism to build the RSE block (Fig. 5). In this block, the overall feature of each channel of the input feature map is calculated as a scalar. Then, the scalar is used as the band weight for multiplication with the feature map. The shortcut is joined to construct an identity map of unweighted features to a high level. In this manner, the spectral value of each channel can be increased or decreased linearly depending on the correlation between them. This imposes constraints on the spectral distribution in the process of image generation and can further improve the color realism of the generated image. We added an RSE block at the first input position of each feature extraction unit (Fig. 4).

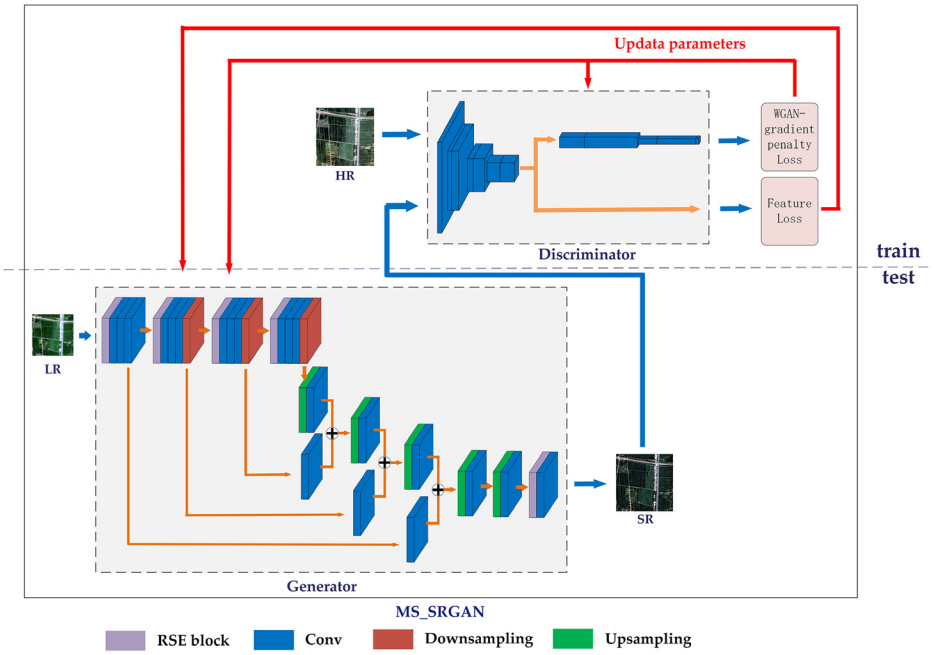

Fig. 4 Basic structure of the MS_SRGAN model (LR: low-resolution image; SR: super-resolution image; HR: high-resolution image; Conv: convolutional layer).

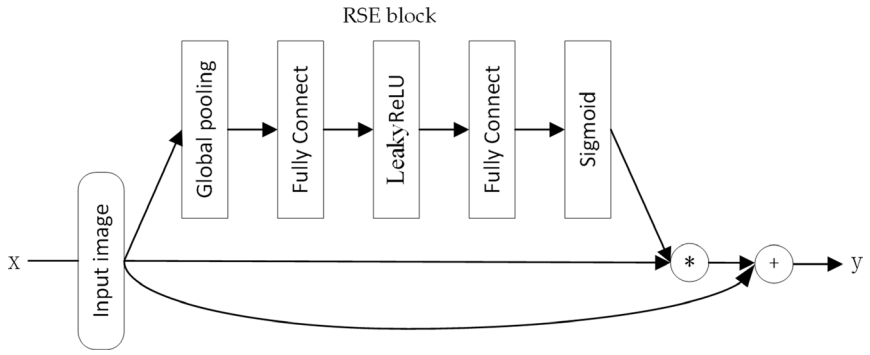

Fig. 5 Schematic of the RSE block.

According to the number of feature extraction units, we added a total of for RSE blocks to the generator.

The overall structure of the generator network, shown in Fig. 4, is composed of two parts. The first part implements feature extraction, and the second part implements scaling enhancement of the feature map.

Feature extraction is carried out step by step and includes four extraction units. Except for the first extraction unit, all others contain one RSE block, three convolutional layers, and downsampling. Each convolutional layer contains two parts: convolution and activation. The size of all convolution kernels is 3×3, and the stride is 1. The number of feature layers doubles with each unit that is passed through. Downsampling is performed by dilated convolution, which is used to reduce the number of rows and columns in a feature map by half. This extraction method fully considers the features of the rich spectral information of remote sensing images, squeezes the redundant features of the multichannel feature map, excites the effective feature of the multichannel feature map, fuses more features in the multichannel space, and defines the functions of different convolutions. At the same time, when the convolutional layer reduces the scale of the feature map, it avoids the occurrence of a large noise or information loss of the channel dimension features.

When the scale of the feature map is enhanced, the recovery unit of each stage and the extraction unit of each step during the feature extraction form a symmetrical structure, and one upsampling layer and two convolutional layers are adopted. This not only restores the scale of the feature map but also establishes a simple feature mapping process for enhancing the low-level feature to the high-level feature, which ensures that the low-level information is not lost due to the reduction of the scale of the feature map. The upsampling layer adopts deconvolution to restore the reduced feature map during feature extraction, and the deconvolution doubles the number of rows and columns of the feature map to restore it to the same size as that of the low-resolution image with a deconvolution kernel of 3×3 and a stride of 2. The convolutional layer is used to adjust the resulting high-level and low-level features to a convolution kernel size of 3×3 and a stride of 1.

Finally, upsampling is carried out to improve the scale of the image. The number of rows and columns are respectively doubled by deconvolution, and the scale of the feature map is raised to be consistent with the size of the high-resolution image. In addition, the spectral distribution of the final feature map is corrected again using the RSE block.

Generator loss is composed of per-pixel, feature, and adversarial losses. Per-pixel loss is calculated for each pixel difference between super-resolution and high-resolution images. The formula for per-pixel loss is as follows:

$$Per-pixel\ loss_{MAE} = \frac{1}{n}\sum_{i=1}^{n} |G(I_i^{LR}) - I_i^{HR}|, \tag{1}$$

where n is the number of batch samples, G represents the generator network, I^{HR} is the high-resolution image, and I^{LR} is the low-resolution image.

Feature loss is realized using the feature map obtained by the convolution of the first seven layers pre-trained by the discriminator. We first train the model without feature loss and then use the first seven convolutional layer parameters of the previously optimized discriminator as the pre-trained network in the subsequent training. The $16 \times 16 \times 256$ size feature map with a larger receptive field is obtained through the pre-trained network to describe the overall feature and control the overall textural structure of the image. The formula for feature loss is as follows:

$$Feature\ loss_{MAE} = \frac{1}{n}\sum_{i=1}^{n} |D'[G(I_i^{LR}) - D'(I_i^{HR})|, \tag{2}$$

where n is the number of batch samples, G represents the generator network, I^{HR} is the high-resolution image, I^{LR} is the low-resolution image, and D' represents the first seven layers of discriminator network.

The generator adversarial loss is part of the discriminator loss. The formula for adversarial loss is as follows:

$$Adversarial\ loss_{Gen} = -\frac{1}{n}\sum_{i=1}^{n} D(G(I_i^{LR})), \tag{3}$$

where n is the number of batch samples, I^{LR} is the low-resolution image, D represents the discriminator network, and G represents the generator network.

The formula for the total loss of the generator is as follows:

$$Loss = \sigma * Per-pixel\ loss_{\mathrm{MAE}} + \beta * Adversarial\ loss_{Gen} + Feature\ loss_{MAE}, \tag{4}$$

where σ and β are the weight coefficients of the pixel loss and adversarial loss, respectively.

3.4 *Discriminator*

The feature extraction structure of the discriminator network is mainly composed of 10 layers of convolution, which can be divided into two types according to their functions. The first type is used to reduce the scale of the feature map; they have a convolution kernel size of 4×4 and a

stride of 2. The second type is used to increase the numbers of the convolution kernels and channels in the feature map; they have a convolution kernel size of 3×3 and a stride of 1. These two types are stacked alternately to compose the feature extraction part of the network. Global average pooling and a 1×1 convolutional layer are selected instead of linear mapping of the full connection layer for the vectorization of feature fitting. In this manner, the feature map can be directly associated with the classification task while reducing model parameters, thereby effectively avoiding discriminator over-fitting.

Many models have shown that the Wasserstein distance is able to effectively avoid the gradient disappearance or gradient explosion during the training of a GAN network. WGAN minimizes the Earth-mover distance by adopting its approximate deformation and truncates the absolute value of the discriminator parameters to no more than a fixed constant 0.01 after each update, which solves the problem of the instability of the GAN during training. Therefore, we chose the Wasserstein distance[49] as the discriminator loss of the model. The formula for the discriminator loss is as follows:

$$Loss = \sup_{\|f\|_l \leq 1} E_{x' \sim P_r}[D(x')] - E_{x \sim P_g}[D(x)], \tag{5}$$

where $\|f\|_l \leq 1$ means that the function is a 1-Lipschitz function, P_g represents the generated image distribution, P_r represents the reference image distribution, x' represents the reference image, x represents the generated image, D represents the discriminator network, and G represents the generator network.

3.5 *Training Steps*

The MS_SRGAN model specific training steps are as follows, with LR representing low-resolution image; SR representing super-resolution image; and HR representing high-resolution image.

1. Generate SR using the LR input generator network.
2. Optimize the discriminator network using the input discriminator of SR and HR. Repeat step (2) K times.
3. Use SR and HR to calculate the generator loss. Use SR and HR to calculate per-pixel loss, and input the pre-trained network to calculate feature loss. Optimize the discriminator network using the SR input to calculate the adversarial loss. Use the weighted sum of the three losses to calculate the generator loss, and optimize the generator network. Repeat steps (1), (2), and (3).

Here, K times is the optimal training for the discriminator, and the overall number of repeated trainings in (1), (2), and (3) is determined by epoch.

3.6 *Experimental Setup*

We selected the Bicubic, EDSR, SRGAN, and ESRGAN models for comparison. Bicubic interpolation is a traditional interpolation method that is the most used super-resolution method in the industry. EDSR is a deep CNN built based on the residual structure. To further enhance the capability of model feature extraction, the batch normalization layer was removed from the model, and the optimized loss target was completely based on the mean absolute error (MAE) index. This method exhibits excellent performance in the super-resolution algorithm of CNN. SRGAN is a classic SRGAN model that adopts standard discriminator loss as GAN loss, whereas generator SRResNet adopts residual structure as the main architecture the construction of the model. ESRGAN is based on SRGAN, but it uses an RRDB as the generator feature extraction block to enhance the feature extraction capability of the model. Its performance is superior to SRGAN in terms of natural image super-resolution. As for the test results of the comparative experimental model, we trained the EDSR, SRGAN, and ESRGAN models from scratch on the dataset in this study and then tested these models (Table 5).

Table 5 Models used in the comparative analyses.

Module name	Model introduction
Bicubic	Traditional bicubic interpolation.
EDSR	A super-resolution CNN composed of the residual structure.
SRGAN	A super-resolution model that processes natural images using SRResNet as a generator.
ESRGAN	Enhanced version of SRGAN. The precision of the result of the ESRGAN is better than that of the SRGAN.
MS_SRGAN	The proposed method.

All experiments in this study were run on a graphic workstation purchased by the laboratory. This workstation is equipped with NVIDIA GeForce GTX TITAN X (Pascal) GPUs with 12 GB of video memory and a Linux Ubuntu 16.04 operating system. In this study, the model was built based on the Pytorch deep learning library, and the coding was implemented in Python language. Additional details of the model are as follows: the number of training images in each batch was 16; the total training epoch was 5,000; the learning rate initialized was $1e-4$; K was 5; and the learning rate decreased to half of the previous rate after every 1000 epochs.

Owing to the limitation of the Wasserstein distance loss, the method described in this paper cannot adopt the optimization method that adds the momentum factor. Therefore, the RMSProp optimization method was adopted for model training; the initial learning rate is 10^{-4}, and the gamma is 0.9. The pre-trained model of the feature loss contains the model parameter that was trained based on the GF1–GF2 data set for the first time, which is saved as a pth weight file, and loaded in each subsequent training. During the training of SRGAN and ESRGAN using the dataset selected in this study, these models also joined the feature loss pre-trained network, and the method was the same as that for MS_SRGAN.

4 Results

4.1 *Evaluation Metrics*

We selected the known performance metrics MAE, structural similarity index measure (SSIM), spectral angle mapper (SAM), and the relative global-dimensional synthesis error (ERGAS) to compare the experimental results of different models.

In our generator, the MAE is a part of the loss. It reflects the level of uncertainty in the image, and it can be used as a performance metric. The formula for MAE is as follows:

$$MAE(x,y)=\frac{1}{W*H}\sum_{i=1}^{W}\sum_{j=1}^{H}|x_{i,j}-y_{i,j}|, \tag{6}$$

where W is the width of the image, H is the height of the image, x is the generated image, and y is the reference image. The ideal result of this metric is 0.

The SSIM[50] compares image distortion in three levels: brightness (mean), contrast (variance), and structure. The formula for SSIM is as follows:

$$SSIM(x,y)=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}, \tag{7}$$

where x is the generated image, y is the reference image, μ_x is the mean value of x, μ_y is the mean value of y, σ_x^2 is the x variance, σ_y^2 is the y variance, σ_{xy} is the x and y covariance,

$c_1 = (0.01 * MAX)^2$, $c_2 = (0.03 * MAX)^2$, and MAX is a constant (65,535).The ideal result of this metric is 1.

The value of MAX is determined by the pixel bit-width of each pixel point in the image. The pixel bit-width of a remote sensing image is different from that of a natural image. Thus, it is meaningless to perform a longitudinal comparison based on the standard of traditional natural images.

The SAM[51] measures the spectral angle between two vectors, and it is used to measure the spectral similarity between the original multispectral data and the reconstructed multispectral data.

$$SAM(v, \hat{v}) = \sin^{-1} \frac{v, \hat{v}}{v_2 * \hat{v}_2}, \tag{8}$$

where v is the pixel vector formed by the reference image and $\hat{v}$ is the vector formed by the generated image. The ideal result of this metric is 0.

The ERGAS[52] provides a global quality evaluation of the generated result and is calculated via Eq. (9).

$$ERGAS = 100 \frac{h}{l} \sqrt{\frac{1}{k} \sum_{i=1}^{k} (RMSE(i)/Mean(i))^2}, \tag{9}$$

$$RMSE(x, y) = \frac{1}{W * H} \sqrt{\sum_{i=1}^{W} \sum_{j=1}^{H} (x_{i,j} - y_{i,j})^2}, \tag{10}$$

where h/l is the ratio between the spatial resolution of the generated image and that of the low-resolution image, k is the number of bands of the generated image, Mean (i) is the mean value of the differences between the i'th band of the reference image and that of the generated image, and RMSE(i) indicates the root-mean-squared error of the i'th band between the reference images y and generated images x. The ideal result of this metric is 0.

4.2 *Super-Resolution Result of GF1*

The test and comparison experiments were carried out on the GF1 test set using the trained MS_SRGAN. Figure 6 contains four groups of images, and each group contains reference HR (GF2 PMS images) and Bicubic (GF1 image bicubic results), EDSR, SRGAN, ESRGAN, and MS_SRGAN generated images in this order from left to right. Each image shows an RGB (red-green-blue) color image, a false color combination NIR-red-green image, and an NIR grayscale image. The a and b groups of images mainly show the results of the crop plantation areas. The c and d groups of images mainly show the results of the areas covered by buildings.

The performance of the five models were compared horizontally using the MAE, SSIM, SAM, and ERGAS metrics of the super-resolution and reference images. As can be seen from Table 6, the results in bold font are the best, and those in italic font represent the second best. The results (Table 6) show that the method presented in this paper performs best in both SSIM and SAM. Its performance is suboptimal in MAE and ERGAS. This is because the loss of the EDSR model is completely based on the MAE index, whereas MS_SRGAN is optimized based on three losses, and its results have the best structural and spectral similarities. Although the SAM metric in Table 6 is very large, the metric calculation program that we wrote strictly followed the metric formula, and we manually verified that the results were correct.

4.3 *Super-Resolution Result of GF6*

Figure 7 shows the test results of the experiment for the GF6 image, which primarily shows the crop plantation areas. It can be seen from Bicubic (GF6 image bicubic results) and reference HR

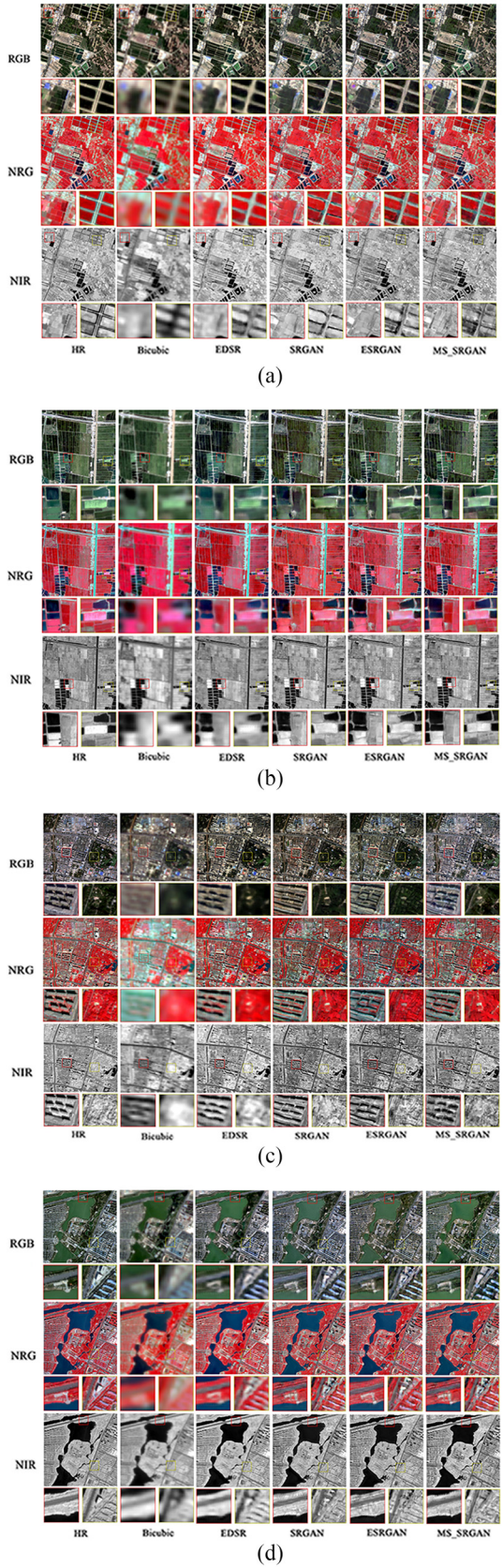

Fig. 6 Reference HR images with corresponding Bicubic, EDSR, SRGAN, ESRGAN, and MS_SRGAN.

Table 6 Performance metrics of the various models.

Metrics	Bicubic	EDSR	SRGAN	ESRGAN	MS_SRGAN
MAEm	107.58	**19.461**	33.589	29.758	*23.125*
SSIMm	0.936	0.994	0.996	*0.997*	**0.998**
SAMm	28.203	7.668	7.534	*5.932*	**3.387**
ERGASm	9.972	**1.332**	3.759	3.947	*2.806*

MAEm: mean MAE; SSIMm: mean SSIM; SAMm: mean SAM; ERGASm: mean ERGAS.

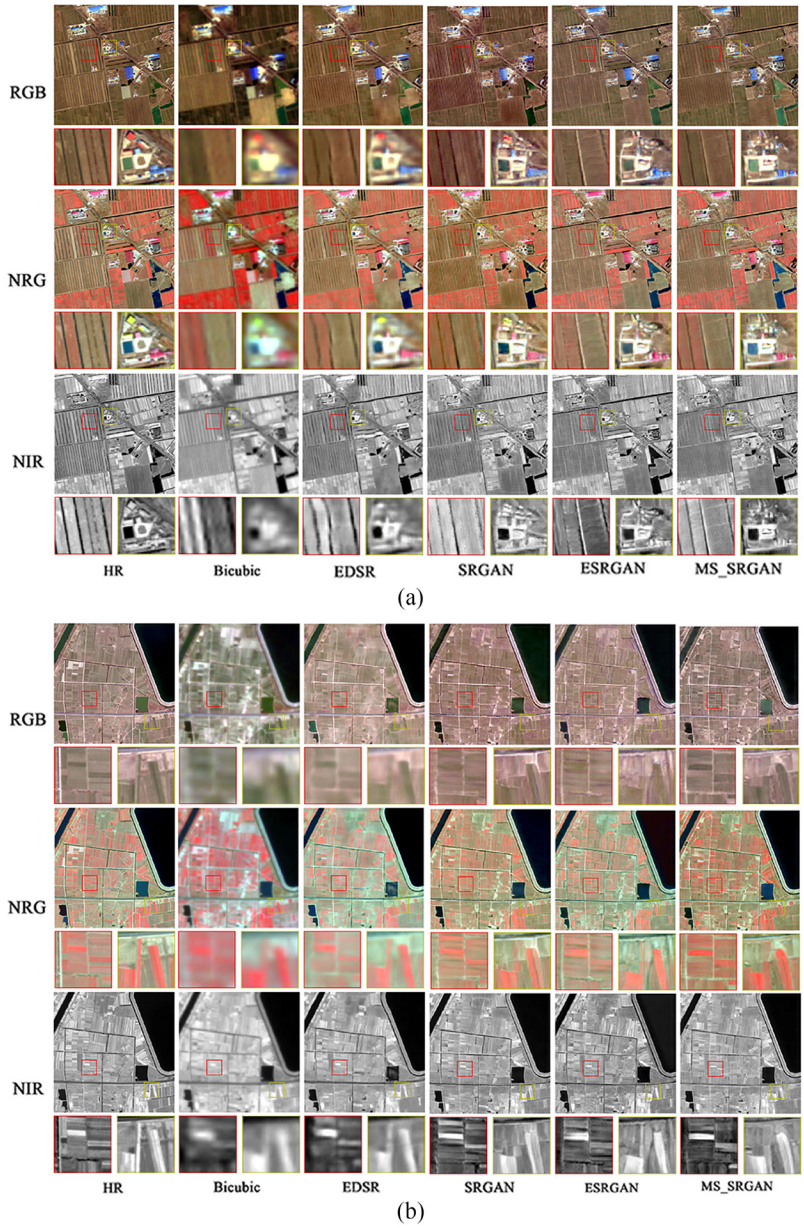

Fig. 7 Reference HR images with corresponding Bicubic, EDSR, SRGAN, ESRGAN, and MS_SRGAN.

Table 7 Performance metrics of the various models.

Metrics	Bicubic	EDSR	SRGAN	ESRGAN	MS_SRGAN
MAEm	147.20	**39.928**	53.811	50.442	*42.571*
SSIMm	0.917	0.993	*0.995*	*0.995*	**0.996**
SAMm	39.635	9.814	9.760	*8.227*	**6.757**
ERGASm	11.541	**4.473**	9.610	8.159	*5.874*

MAEm: mean MAE; SSIMm: mean SSIM; SAMm: mean SAM; ERGASm: mean ERGAS.

(GF2 PMS images) that there is a substantial time difference between the low-resolution image and the high-resolution image.

The performance of the five models were compared horizontally using the MAE, SSIM, SAM, and ERGAS metrics of the super-resolution and reference images. In Table 7, the bold font represents the best and the italic font represents the second-best results.

5 Discussion

A comparison of the experimental results in which, owing to the large difference between the low-resolution and the reference images, the Bicubic method that is based on interpolation can only show the basic characteristics of the low-resolution image (GF1 image), the result is blurred, and the sharpening effect is poor. The result of the EDSR model that is based on the CNN is close to that of the reference image in terms of high-spatial-resolution to some extent, but it still has the problem of blurring. In contrast, the three models that are based on the GAN have a better sharpening effect and clarity; however, details of "artifacts" are provided only to a certain extent by SRGAN and ESRGAN. Overall, the MS_SRGAN method provides the most realistic high-resolution images.

This can be further illustrated through the point distribution histograms in Fig. 8. We used the GF2 image as a reference; the goal is to get a high-resolution image with a spatial resolution close to that of the GF2 image by the model. The closer the result is to the distribution of the GF2 image, this stronger the sense of reality is and the higher the accuracy of the image obtained by the model is. Among all of the models, MS_SRGAN has the maximum similarity in terms of the pixel value distribution range and pixel value curve trend of the reference image. Therefore, the spatial details and texture information of MS_SRGAN are closer to the reference image, which can further improve the accuracy of the results of subsequent applications.

5.1 *Influence of Dataset Production and Criteria*

Two patterns for the establishment of a remote sensing image reconstruction algorithm dataset exist. The simple pattern is mainly aimed at the visible spectra, and the remote sensing data format is compressed into the RGB color mode for super-resolution. The advantage of this method is that, after the band value is compressed, the data complexity is reduced, and the image can be represented to a certain extent. However, the accuracy of the compressed data is greatly reduced. The other approach, which was employed in this study, is based on adding other spectra while retaining the original format of the remote sensing images.

In this study, GF1 or GF6 WFV images were used as low-resolution images, and GF2 PMS images were used as high-resolution images, while the original format of remote sensing image data was retained. This method captures the high-frequency details of high-resolution images to the maximum extent and avoids the loss of information in the process of image compression. However, we ran into problems in the experiment. As can be seen from the point distribution histogram given in Sec. 3.1, the pixel value distribution of remote sensing images significantly differs, which leads to a phenomenon similar to the pixel value distribution shift in EDSR,

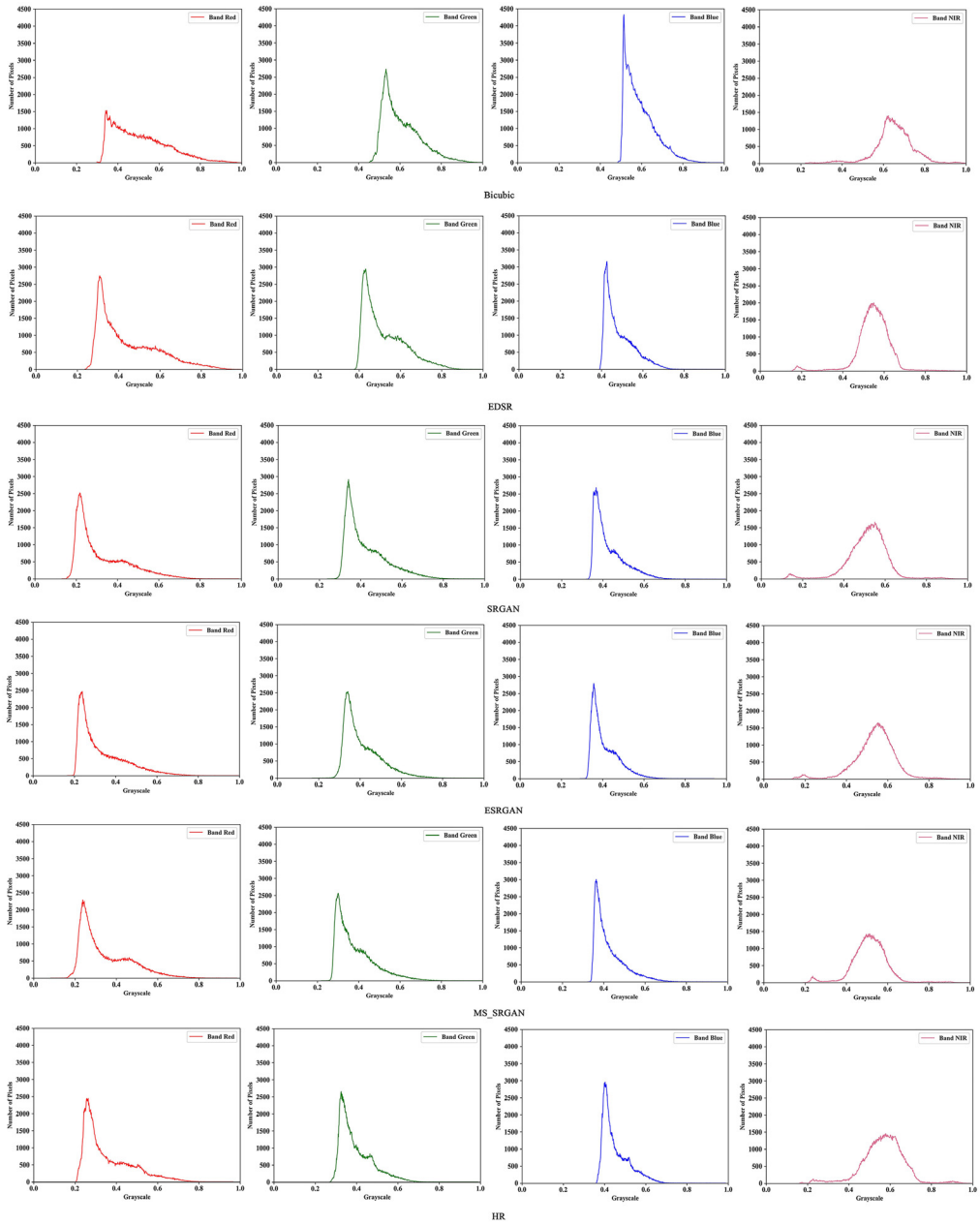

Fig. 8 Point distribution histograms of Bicubic, SRGAN, ESRGAN, MS_SRGAN, and HR.

SRGAN, and ESRGAN, as shown in Fig. 8. To overcome this problem, a model with a strong ability of fitting and spectral information correction is needed.

In addition, the long time interval between low-resolution images and reference images is an obstacle to model training. As the main surface features of the Shandong dataset are crops that were still growing in March, the surface features of the low-resolution images were considerably different from those of the reference images, which further increased the difficulty of model training.

Through repeated experiments and data production, several factors affecting the experimental results, such as the time intervals of the low-resolution and reference images, pixel value distributions of the low-resolution and reference images, and image cloud interference, were identified. Selecting images with a small time interval and accurate and consistent surface features are important criteria for dataset construction.

5.2 *Influence of Different Generator Network Structures*

The results of the comparative analyses (Fig. 8) indicate that MS_SRGAN exhibited a higher performance than the other models in achieving super-resolution, implying that the model's improvement of multispectral images is effective. The RSE block plays an important role in the model because this attention mechanism can effectively highlight more prominent features of the channel and realize the model's ability to correct the spectral information.

To choose the upsampling layer, we conducted tests on both the deconvolution and sub-pixel convolution. The test results show that deconvolution can generate super-resolution results more quickly and efficiently. Finer results can be obtained with additional training of sub-pixel convolution, but these finer results are often different from the real images. Therefore, we chose deconvolution as the upsampling layer because it is more efficient and has a better performance.

5.3 *Influence of Surface Feature Type on Super-Resolution*

In Fig. 6 show that the results of buildings by MS_SRGAN are poorer than those of crops, which can be attributed to the images in the GF1-GF2 dataset only contain a small part of architectural features, with a relatively sparse distribution of buildings. In addition, some buildings contain only a few pixel blocks in the low-resolution image, causing difficulty in obtaining high-frequency textural information for the model. In contrast, crop coverage was large, and hence, the results were better than those for buildings.

The NIR band is included in the red band, and it is mainly used to detect the existence of O-H (O: oxygen and H: hydrogen), N-H (N: nitrogen), and C-H (C: carbon) bonds in substances. The NIR band is often used for the monitoring and analysis of plants because plants mostly contain these chemical bonds. The test results given in Fig. 7 show that the reconstruction result was the poorest for the NIR band of crops among all bands. This is because the crops contain large quantities of the O-H, N-H, and C-H chemical bonds, but there are differences in the content of the chemical bonds between individual plants. Therefore, the high-frequency textural details of the crops in the NIR band are complex, and it is more difficult to achieve super-resolution.

6 Conclusions

This study proposed a new method, MS_SRGAN, for obtaining large coverage and high-resolution multispectral images. This method took the GF2 PMS multispectral image as the high-resolution image and carried out super-resolution for the GF1 WFV multispectral image. The advantages of MS_SRGAN in the super-resolution reconstruction of multispectral images were confirmed through experimental comparison with the Bicubic, EDSR, SRGAN, and ESRGAN methods. This paper discussed the influence of dataset production and criteria, different generator network structures, and surface feature type on super-resolution, and it explored the advantages and disadvantages of the new MS_SRGAN method in detail.

To retain rich spectral information of remote sensing data, this study's training dataset retains the original data format. In the experiment, we found that there is a problem of inconsistent spectral distribution between the reference image and low-resolution image. To solve this problem, this method joins an RSE block to construct the generator network and adds the Wasserstein distance as the discriminator loss to perform super-resolution of multispectral images. By training different data, a set of criteria and methods for creating datasets with different remote sensing images as references were determined.

However, we found that the super-resolution of NIR bands of crops and complex surface features such as small villages and mountains was not ideal because it is difficult to provide high-frequency textural details on low-resolution images and the generalization ability of the model needs improvement.

In future studies, we hope to add more types of spectral images to improve the accuracy of the image, improve the model architecture and loss function to enhance the generalizability of the model, and find a more effective method for evaluating the super-resolution of remote sensing image.

Acknowledgments

This research was funded by the Key Research and Development Program of Ningxia (Award no: 2019BEH03008); the National Key Research and Development Program of China (Award No. 2018YFC1506500); the Applied Foundation of Qinghai(Award No. 2021-ZJ-739); the Science Foundation of Shandong (Award No. ZR2020MF130); the arid meteorological science research fund project by the Key Open Laboratory of Arid Climate Change and Disaster Reduction of CMA (Award No. IAM201801). We thank the Supercomputing Center in Shandong Agricultural University for technical support.

References

1. M. G. Hu, J. F. Wang, and C. M. Chen, "Mixed-pixel decomposition and super-resolution reconstruction of RS image," *Prog. Geogr.* **29**(6), 747–756 (2010).
2. D. Yan, Z. Li, and Y. Xia, "Remote sensing image super-resolution: challenges and approaches," in *IEEE Int. Conf. Digital Signal Process.*, Singapore, pp. 196–200 (2015).
3. J. L. Harris, "Diffraction and resolving power," *J. Opt. Soc. Am.* **54**(7), 931–933 (1964).
4. R. Tsai and T. Huang, "Multi-frame image restoration and registration," *Adv. Comput.; Vis. Image Process* **1**, 317–339 (1984).
5. X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Trans. Image Process.* **17**(6), 887–896 (2008).
6. W. S. Dong, L. Zhang, and G. M. Shi, "Nonlocal back-projection for adaptive image enlargement," in *16th IEEE. Int. Conf. Image Process. (ICIP)*, Cairo, pp. 349–352 (2009).
7. Y. W. Tai et al., "Super resolution using edge prior and single image detail synthesis," in *IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recognit.*, San Francisco, California, pp. 2400–2407 (2010).
8. M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and under sampled measured images," *IEEE Trans. Image Process.* **6**(12), 1646–1658 (1997).
9. X. X. Wang, *Study on Super-Resolution Image Restoration Based on Dictionary Learning and Sparse Representation*, Harbin University of Science and Technology (2014).
10. J. C. Yang, J. Wright, and T. S. Huang, "Image super-resolution as sparse representation of raw image patches," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Anchorage, AK, pp. 1–8 (2008).
11. R. Zeyde, M. Elad, and M. Protter, "On single image scale-up using sparse-representations," *Lect. Notes Comput. Sci.* **6920**, 711–730 (2010).
12. J. Yang et al., "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Process.* **21**(8), 3467–3478 (2012).
13. R. Timofte, V. De Smet, and L. Van Gool, "Anchored neighborhood regression for fast example-based super-resolution," in *IEEE Int. Conf. Comput. Vision*, Sydney, NSW, pp. 1920–1927 (2013).
14. R. Timofte, V. De Smet, and L. Van Gool, "A+: adjusted anchored neighborhood regression for fast super-resolution," in *Proc. Asian Conf. Comput. Vision*, Cham, Springer, pp. 111–126 (2014).
15. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. Learn. Represent., ICLR 2015—Conf. Track Proc* (2015).
16. K. He et al., "Deep residual learning for image recognition," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Las Vegas, Nevada, pp. 770–778 (2016).
17. G. Huang et al, "Densely connected convolutional networks," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Honolulu, Hawaii, pp. 2261–2269 (2017).
18. M. Lin, Q. Chen, and S. Yan, "Network in network," https://arxiv.org/pdf/1312.4400v2.pdf (2014).
19. C. Szegedy, W. Liu, and Y. Jia, "Going deeper with convolutions," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Boston, Massachusetts, pp. 1–9 (2015).
20. C. Dong, et al., "Learning a deep convolutional network for image super-resolution," *Lect. Notes Comput. Sci.* **8692**, 184–199 (2014).

21. J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Las Vegas, Nevada, pp. 1646–1654 (2016).
22. X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using convolutional auto-encoders with symmetric skip connections," *Proc. Advances in Neural Inf. Process. Syst.*, https://arxiv.org/pdf/1606.08921.pdf (2016).
23. J. Kim, J. K. Lee, and K. M. Lee, "Deeply-recursive convolutional network for image super-resolution," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Las Vegas, Nevada, pp. 1637–1645 (2016).
24. W. S. Lai et al., "Fast and accurate image super-resolution with deep Laplacian pyramid networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **41**(11), 2599–2613 (2019).
25. T. Tong et al., "Image super-resolution using dense skip connections," in *IEEE Int. Conf. Comput. Vision*, Venice, pp. 4809–4817 (2017).
26. B. Lim et al., "Enhanced deep residual networks for single image super-resolution," in *IEEE Conf. Comput. Vision and Pattern Recognit. Workshops*, Honolulu, Hawaii, pp. 1132–1140 (2017).
27. Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," *Lect. Notes Comput. Sci.* **11211**, 294–310 (2018).
28. C. Dong, C. C. Loy, and X. Tang, "Accelerating the super-resolution convolutional neural network," *Lect. Notes Comput. Sci.* **9906**, 391–407 (2016).
29. W. Z. Shi et al., "Real-time single image and video super-resolution using an efficient subpixel convolutional neural network," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Las Vegas, Nevada, pp. 1874–1883 (2016).
30. J. I. Goodfellow et al, "Generative adversarial networks," in NIPS, https://arxiv.org/pdf/1406.2661.pdf (2014).
31. J. Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE Int. Conf. Comput. Vision*, Venice, pp. 2242–2251 (2017).
32. X. X. Qu et al., "Perceptual-DualGAN: perceptual losses for image to image translation with generative adversarial nets," in *Int. Joint Conf. Neural Networks*, Rio de Janeiro, pp. 1–8 (2018).
33. K. Lata, M. Dave, and K. N. Nishanth, "Image-to-image translation using generative adversarial network," in *3rd Int. Conf. Electron. Commun. and Aerosp. Technol.*, Coimbatore, India, pp. 186–189 (2019).
34. S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graphics* **36**(4), 1–14 (2017).
35. Y. Li et al., "Generative face completion," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Honolulu, Hawaii, pp. 5892–5900 (2017).
36. C. X. Zheng, T. Cham, and J. F. Cai, "Pluralistic image completion," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Long Beach, California, pp. 1438–1447 (2019).
37. J. W. Chen et al., "Image blind denoising with generative adversarial network based noise modeling," in *IEEE/CVF Conf. Comput. Vision and Pattern Recognit.*, Salt Lake City, Utah, pp. 3155–3164 (2018).
38. A. Alsaiari et al., "Image denoising using a generative adversarial network," in *IEEE 2nd Int. Conf. Inf. and Comput. Technol.*, Kahului, Hawaii, pp. 126–132 (2019).
39. Y. Zhong et al., "A generative adversarial network for image denoising," *Multimedia Tools Appl.* **79**(2), 16517–16529 (2020).
40. C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conf. Comput. Vision and Pattern Recognit.*, Honolulu, Hawaii, pp. 105–114 (2017).
41. J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Lect. Notes Comput. Sci.* **9906**, 694–711 (2016).
42. X. Wang et al., "ESRGAN: enhanced super-resolution generative adversarial networks," *Lect. Notes Comput. Sci.* **11133**, 63–79 (2019).
43. Y. Saquil, K. Kim, and P. Hall, "Ranking CGANs. Subjective control over semantic image attributes," https://arxiv.org/pdf/1804.04082.pdf (2018).

44. W. Zhang et al., "RankSRGAN: generative adversarial networks with ranker for image super-resolution," in *IEEE/CVF Int. Conf. Comput. Vision*, Seoul, pp. 3096–3105 (2019).
45. D. Pouliot et al., "Landsat super-resolution enhancement using convolution neural networks and Sentinel-2 for training," *Remote Sens.* **10**(3), 394 (2018).
46. Z. F. Shao et al., "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sens. Environ.* **235**, 111425 (2019).
47. J. Zhang, S. Liu, and Y. Peng, "Satellite image super-resolution based on progressive residual deep neural network," *J. Appl. Remote Sens.* **14**(3), 032610 (2020)
48. M. Wen et al., "Super-resolution of remote sensing images based on transferred generative adversarial network," in *IGARSS 2018-2018 IEEE Int. Geosci. and Remote Sens. Symp.*, IEEE (2018).
49. M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," ArXiv abs/1701.07875, 2017, http://proceedings.mlr.press/v70/arjovsky17a/arjovsky17a.pdf.
50. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
51. 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**, 691–699 (1997).
52. L. Alparone et al., "Comparison of pan-sharpening algorithms: outcome of the 2006 GRS-S data fusion contest.," *IEEE Trans. Geosci. Remote Sens.* **45**, 3012–3021 (2007).

Ziyun Zhang has a master's degree in information science and engineering, Shandong Agricultural University. His main research interests are remote sensing image super resolution and remote sensing image segmentation. At present, he is mainly engaged in the research of remote sensing technology for agriculture and environment.

Chengming Zhang is currently a professor working at the College of Information Science and Engineering of Shandong Agricultural University. His main research areas are remote sensing and geographic information system in land use monitoring and evaluation, presided over a number of agricultural remote sensing projects by Ministry of Science and Technology and Shandong Province. Currently, he is mainly engaged in the research of remote sensing technology in agriculture and environment.

Biographies of the other authors are not available.