SWCGAN: Generative Adversarial Network Combining Swin Transformer and CNN for Remote Sensing Image Super-Resolution

Jingzhi Tu, Gang Mei, Zhengjing Ma, and Francesco Piccialli

***Abstract*—Easy and efficient acquisition of high-resolution remote sensing images is of importance in geographic information systems. Previously, deep neural networks composed of convolutional layers have achieved impressive progress in super-resolution reconstruction. However, the inherent problems of the convolutional layer, including the difficulty of modeling the long-range dependency, limit the performance of these networks on super-resolution reconstruction. To address the abovementioned problems, we propose a generative adversarial network (GAN) by combining the advantages of the swin transformer and convolutional layers, called SWCGAN. It is different from the previous super-resolution models, which are composed of pure convolutional blocks. The essential idea behind the proposed method is to generate high-resolution images by a generator network with a hybrid of convolutional and swin transformer layers and then to use a pure swin transformer discriminator network for adversarial training. In the proposed method, first, we employ a convolutional layer for shallow feature extraction that can be adapted to flexible input sizes; second, we further propose the residual dense swin transformer block to extract deep features for upsampling to generate high-resolution images; and third, we use a simplified swin transformer as the discriminator for adversarial training. To evaluate the performance of the proposed method, we compare the proposed method with other state-of-the-art methods by utilizing the UCMerced benchmark dataset, and we apply the proposed method to real-world remote sensing images. The results demonstrate that the reconstruction performance of the proposed method outperforms other state-of-the-art methods in most metrics.**

***Index Terms*—Convolutional layers, generative adversarial network (GAN), remote sensing images, super-resolution reconstruction, swin transformer.**

I. Introduction

CURRENTLY, remote sensing plays an important role in various fields [1]–[4]. As a critical component of remote sensing, optical satellite remote sensing with high spatial resolution provides observation targets with clear spatial texture information, which can present the essential features of the landscape. However, an increase in the pixel density of the sensor will significantly increase the hardware cost of obtaining optical remote sensing images. To acquire high-resolution remote sensing images more conveniently and cost-effectively, super-resolution reconstruction techniques that recover high-resolution images according to low-resolution images have received much attention.

The essential idea behind image super-resolution is to learn the prior knowledge from the image data and then use it to recover the lost details of the low-resolution images. It is worthwhile to mention that some early methods, such as sparse reconstruction-based methods [5], degenerate models [6], and interpolation [7], have achieved success in high-resolution images by acquiring a priori knowledge under the limitation of a lack of learning ability. In recent years, deep learning methods have achieved great success in various fields, and naturally, super-resolution reconstruction methods with powerful learning capabilities based on deep learning surpass traditional methods in terms of performance. Convolutional neural networks (CNNs) have long been the "standard answer" to image processing tasks. A super-resolution CNN, the first CNN with a high learning ability for super-resolution reconstruction, was proposed by Dong *et al.* [8]. Then, to further boost the performance of super-resolution reconstruction using CNNs, various new techniques, including residual dense blocks [9], [10], residual learning [11], [12], and recursive blocks [13], are also introduced.

Moreover, with generative adversarial networks (GANs) being originally proposed and receiving great attention [14], [15], GANs using convolution modules have also been proven to have impressive performance on the super-resolution reconstruction task. For example, SRGAN [16] first introduced a GAN architecture for the super-resolution reconstruction task and proposed a perceptual similarity-based loss function. CDGAN [17] has improved the discriminator for GAN, where both generated image and its high-resolution ground truth are input to the discriminator for better discrimination. EEGAN [18] reduces the interference of noise in the super-resolution reconstruction of satellite images by purifying the noise-contaminated components with mask processing. Although CNNs with powerful learning capabilities offer a significant performance improvement over traditional

Manuscript received 22 November 2021; revised 13 February 2022, 9 June 2022, and 2 July 2022; accepted 9 July 2022. Date of publication 13 July 2022; date of current version 22 July 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 11602235 and in part by the Fundamental Research Funds for China Central Universities under Grant 2652021053. *(Corresponding author: Gang Mei.)*

Jingzhi Tu, Gang Mei, and Zhengjing Ma are with the School of Engineering and Technology, China University of Geosciences, Beijing 100083, China (e-mail: tujingzhi@cugb.edu.cn; gang.mei@cugb.edu.cn; zhengjing.ma@cugb.edu.cn).

Francesco Piccialli is with the Department of Mathematics and Applications "R. Caccioppoli", University of Naples Federico II, 80138 Napoli, Italy (e-mail: francesco.piccialli@unina.it).

Digital Object Identifier 10.1109/JSTARS.2022.3190322

methods, they cannot escape from the basic problems that originate from the basic convolutional layer, i.e., the convolutional layer based on the principle of local processing has difficulty, or is even ineffective, for the capture of long-range dependencies.

To solve the abovementioned problem, a self-attention mechanism derived from transformer [19]–[21] was used as an alternative to CNNs that capture global interactions between contexts and shows an excellent performance on several visual tasks. However, the networks designed based on the transformer block often have a number of parameters that exceed those of general convolutional networks. On the other hand, the transformer for image recovery typically segments the input image into patches [22]–[24], which can introduce boundary artifacts around each patch.

Recently, swin transformer [25], a self-attention network that overcomes the abovementioned shortcomings of the transformer, has shown great potential in the field of computational vision. It is capable of both processing large-size images without dividing the images into patches and learning long-range dependencies as a transformer due to the shifted window scheme. Furthermore, it has less computational cost than the transformer. Swin transformer has reached state-of-the-art in image classification and semantic segmentation tasks. However, its application and research in image super-resolution, especially in remote sensing images, is still relatively rare.

Remote sensing images contain more information than natural images, and the pixels of remote sensing images are correlated with each other. CNNs have difficulties in acquiring global information and long-range dependencies between pixels for remote sensing images. Moreover, since the size of real remote sensing images tends to be larger than that of natural images and the complexity of the self-attention network is high, the pure self-attention network is prone to memory bottlenecks if the remote sensing images are used directly as input.

In this article, to adapt the characteristics of remote sensing images, we first introduce the shifted window self-attention mechanism from the swin transformer into the super-resolution research of remote sensing images and propose a GAN that combines the advantages of the swin transformer and convolutional layers, namely, SWCGAN. Specifically, in the proposed SWCGAN,

1) we employ a convolutional layer for shallow feature extraction that can be adapted to flexible input sizes;
2) we propose the residual dense swin transformer block (RDSTB) by drawing on the characteristics of DenseNet [9] to build the depth feature extraction module of the generator, which is used to obtain the deep features for upsampling to generate high-resolution images;
3) we simplify the original swin transformer and use it as a discriminator.

To demonstrate the performance of the proposed SWCGAN, the UCMerced dataset is utilized for training and validation of the proposed method, and the performance of the proposed method outperforms other state-of-the-art methods in most metrics. Moreover, the proposed method is applied to a real-world remote sensing image to verify the effectiveness and applicability of the proposed method.

Our contributions are as follows.

1) We propose a GAN with a hybrid of convolutional and swin transformer layers for the super-resolution reconstruction of remote sensing images to consider the features of large size, large information, and strong correlation between pixels of remote sensing images for super-resolution reconstruction.
2) We further propose a depth feature extraction block, namely, RDSTB, which can extract deep image features efficiently by stacking multiple blocks. As a feature extraction block of images, the proposed RDSTB can also be used in other image processing tasks in the future.
3) We evaluate the proposed method using the UCMerced benchmark dataset and real-world remote sensing images from a high-resolution satellite.

The rest of this article is organized as follows. In Section II, the proposed method, including the network architecture, the loss functions, and the shifted window self-attention mechanism, are described in detail. Section III presents the experimental results and performance evaluations, and the proposed method is applied to a real-world remote sensing image. In Section IV, we present a discussion. Finally, Section V concludes this article.

II. Methods

In this section, we will first briefly introduce the essential idea behind the proposed method, i.e., the idea of the SWCGAN for super-resolution. Specifically, for the proposed method, we will describe the generator network, the discriminator network, and the loss function. Furthermore, the advanced shifted window self-attention mechanism is summarized.

A. Overview of SWCGAN

In this article, we proposed a GAN by combining the advantages of the swin transformer and convolutional layers for super-resolution. The workflow of the proposed SWCGAN for super-resolution is illustrated in Fig. 1. A typical GAN model consists of two parts, a generator G and a discriminator D. As shown in Fig. 1(a), for the generator G, the input is the low-resolution image, and then, the features of the input low-resolution image is extracted to obtain the feature maps using the feature extraction module. To obtain the generated high-resolution image, the extracted feature maps will be upsampled by the upsampler (in this article, 4x upsampling is used). For the discriminator D, the input is the generated high-resolution image, and the generated high-resolution image also undergoes a feature extraction step to obtain its deep features, where the sizes of feature maps gradually decrease, unlike in the generator. Finally, a linear layer is classified dichotomously by the output 0 or 1. Naturally, in an image super-resolution GAN, G is used to generate a fake high-resolution image by reducing the difference between the fake high-resolution image and the real high-resolution image, and D is used to distinguish the real high-resolution image from the generated image in training. G and D compete with each other in the training process in such a way that the data distribution of the generated images is gradually close to the real distribution [see Fig. 1(b)].

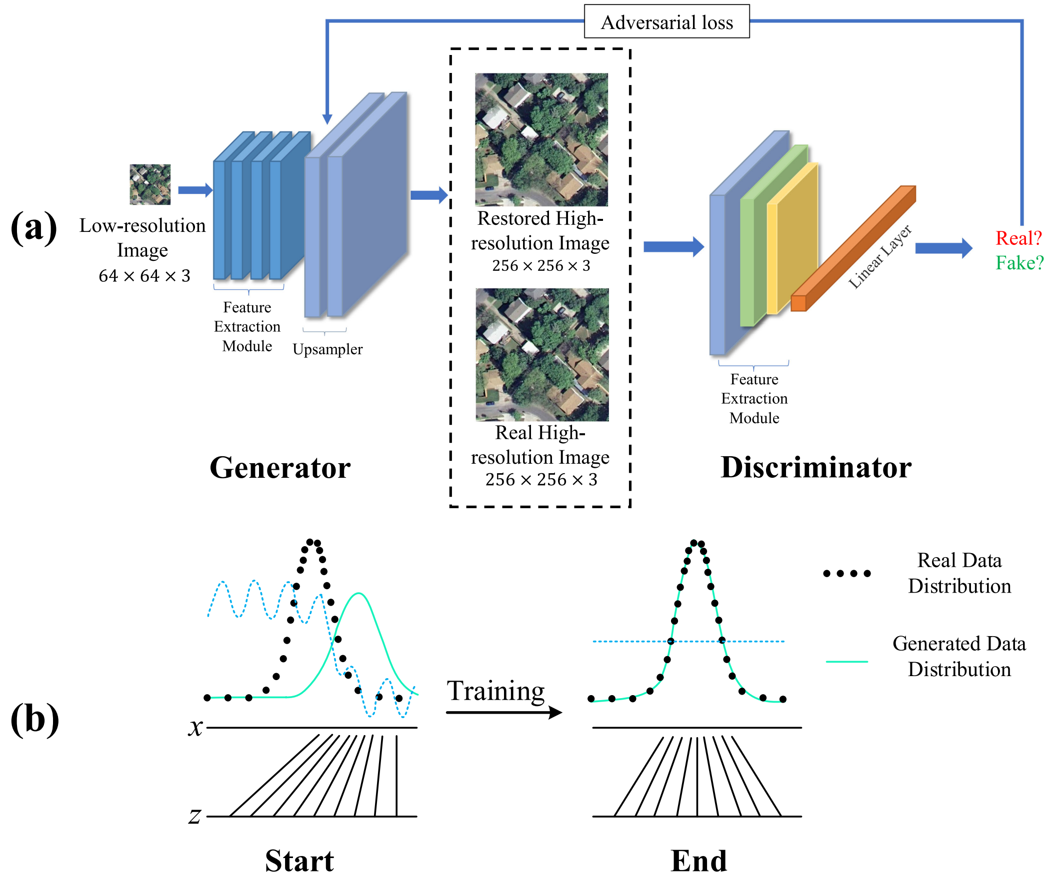


Fig. 1. Illustration of GAN for the super-resolution reconstruction task. (a) Network framework. (b) Data distribution.

Previously, convolutional blocks were commonly used to form GANs for super-resolution reconstruction. However, the convolutional layers that compose CNNs limit the performance of CNNs in the super-resolution reconstruction task due to its inherent problems, especially the inability to simulate long-term dependencies. To address the abovementioned issue, we propose a GAN with a hybrid of convolutional and swin transformer layers named SWCGAN for the super-resolution reconstruction task, where we introduce swin transformer layers with convolutional layers to form RDSTB by dense connection and residual structure to extract image features.

B. Generator Network in SWCGAN

As shown in Figs. 1(a) and 2(a), the generator can be further divided into the following three modules:

1) shallow feature extraction module;
2) deep feature extraction module;
3) upsampling module.

In this shallow feature extraction module, a convolution layer $\mathrm{Conv}(c, n_f, k, s, p)$ is used to extract shallow features, where c represents the filter channels, n_f represents the number of filters, k represents the kernel size, s represents the stride, and p represents the padding. The advantage of the convolution layer in shallow image processing is that it can flexibly set the input image size. Thus, k, s, and p should be set to 3, 1, and 1, which ensures that the channel of the input image is transformed from c to n_f without changing the size of the input image ($H \times W \times 3$ to $H \times W \times n_f$).

After extracting shallow features, the $H \times W \times n_f$ feature maps are imported into the deep feature extraction module, which is composed of the L layer of the proposed RDSTB (in this article, we set $L = 4$). As shown in Fig. 2(b), in this proposed RDSTB, we first introduce a swin transformer layer and combine it with a convolutional layer, which ensures the learning of local information and the modeling of long-range dependency with shifted window self-attention. Then, inspired by DenseNet, we connect these residual blocks in a densely connected scheme (N is set to 6), which keeps us building a complex and deep model. Finally, we keep the dimensions of the input and output of RDSTB constant by a convolution layer.

As shown in Fig. 3, the nearest neighbor interpolation method [26], [27] combined with a convolutional layer is applied to upsample the feature map after extracting deep feature, and its size changes from $H \times W$ to $2H \times 2W$. We utilized this upsampling operation twice to achieve a 4x upsampling effect. There have been some investigations showing that this upsampling operation is effective in reducing the noise of the generated high-resolution images compared to the upsampler composed of convolution and PixelShuffle [28]. Moreover, to comprehensively utilize the extracted features, we aggregate

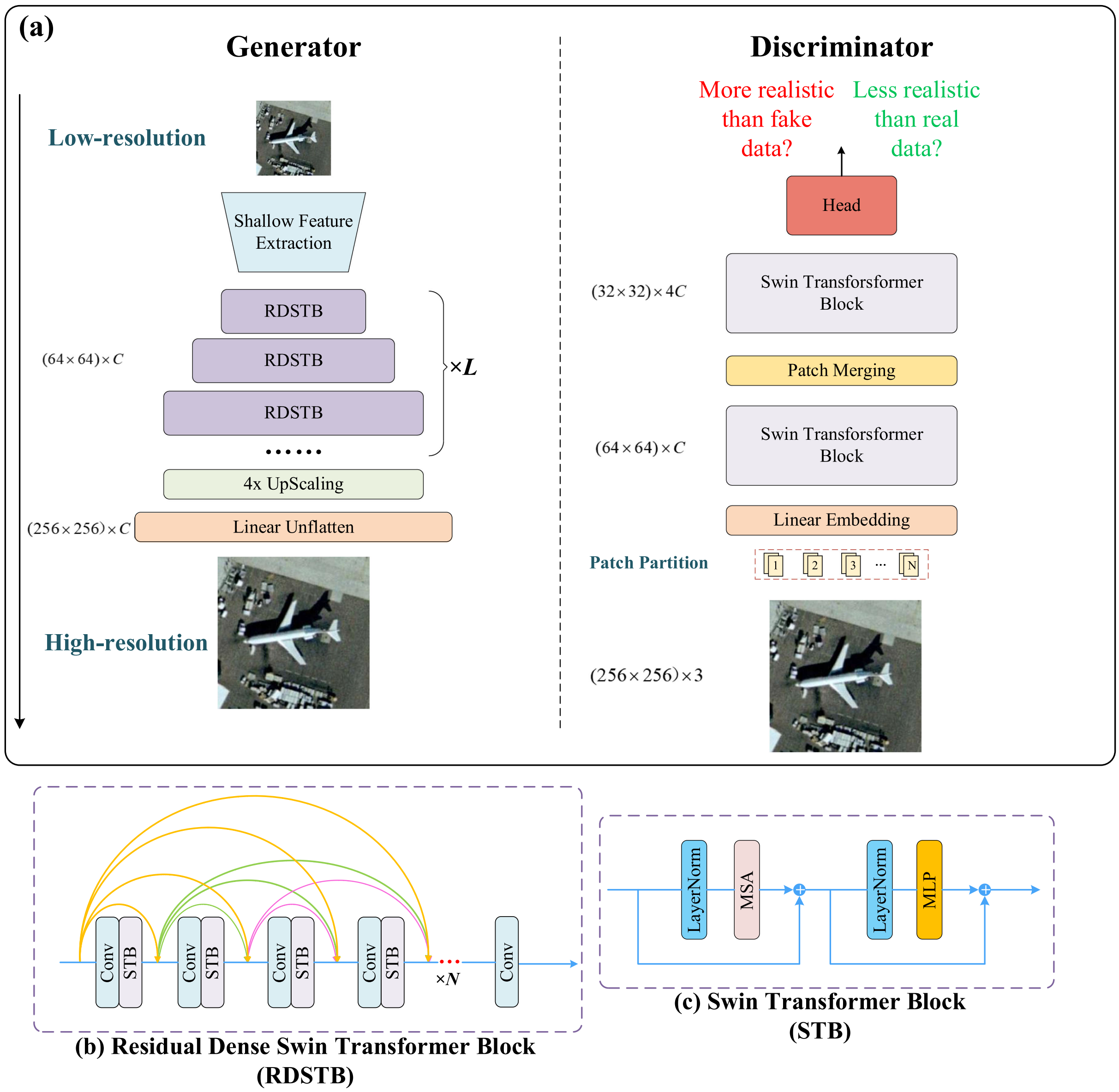


Fig. 2. Architecture of the proposed SWCGAN. (a) Network architecture of generator and discriminator. (b) RDSTB. (c) Swin transformer block.

shallow features and deep features into the upsampler

$$I_{\mathrm{HR}} = H_{\mathrm{up}}(F_0 + F_D) \tag{1}$$

where I_{HR} is the generated high-resolution image, H_{up} is the function of the upsampler, F_0 is the shallow features, and F_D is the deep features.

C. Discriminator Network in SWCGAN

For the discriminator, the simplified swin transformer is used to conduct the dichotomous classification task. As shown in Figs. 2(a) and 4, in the original swin transformer, the feature extraction is a total of 4 stages, and the dimensions of the input data are transformed from $H \times W \times 3$ to the feature maps $H/32 \times W/32 \times 8\,C$ after extracting features. In our simplified swin transformer (discriminator), we employ only the first two stages of the original swin transformer as the feature extraction module of the discriminator, and the dimensions of the input data are transformed from $H \times W \times 3$ to the feature maps $H/8 \times W/8 \times 2\,C$ after extracting features. Then, the feature maps $H/8 \times W/8 \times 2\,C$ from the feature extraction module are used directly for classification by a linear layer. It is worthwhile to note that instead of the standard discriminator, we use the relativistic average discriminator, which estimates the probability that the real image is relatively more realistic than the fake image. Its concrete implementation is described in the loss function.

D. Use of Swin Transformer in SWCGAN

To address the challenges involved in the application of CNNs and transformers to the image field, the swin transformer proposed a shifted windows operation that included nonoverlapping local windows and overlapping cross-window connections to restrict the attentional computation to a single window, which

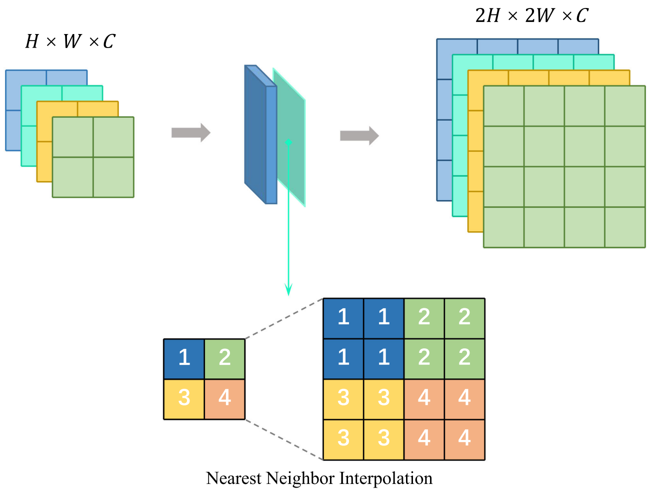


Fig. 3. Illustration of the upsampler in SWCGAN.

can allow the model to enjoy the advantages of CNN convolutional operations on the one hand and to save computational effort on the other hand.

As shown in Fig. 4, the whole model adopts a hierarchical design, where the model contains a total of four stages, each of which reduces the resolution of the input feature map and expands the receptive field layer by layer like a CNN. First, compared with vision transformer [23], [29], patch partitioning becomes an optional operation due to the window self-attention mechanism, which greatly increases the flexibility of the model. Second, there is a patch merging operation before performing swin transformer block, which serves to perform downsampling. It is used to reduce the resolution, adjust the number of channels, and save computational effort. In patch merging, the features of each group of 2×2 neighboring patches are concatenated together as a whole tensor, and a linear layer is applied to these $4\,C$-dimensional concatenated features, reducing its output dimension to $2\,C$. Third, the architecture of the STB is similar to that of the transformer block, except that the standard multihead self-attention module in the transformer block is replaced with a module built on the shifted windows, and the other layers remain the same.

To solve the problem of high computational complexity caused by the global-based calculation of attention in the traditional transformer, the swin transformer reduces the complexity of the algorithm by the window attention, limiting the computation of attention to each window. The equation for computing self-attention is as follows:

$$A(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}} + B\right)V \tag{2}$$

where Q, K, and $V \in \mathbb{R}^{M^2 \times d}$ indicate the query, key, and value matrices, $B \in \mathbb{R}^{M^2 \times M^2}$ indicates a relative position bias; d represents the $\frac{Q}{K}$ dimension, and M^2 represents the number of patches in a window ((M, M) is the window size). In the computation of self-attention, the relative position is critical.

The abovementioned window self-attention is calculated for each window. To allow different windows to interact with one another, a unique method called shifted window is adopted in swin transformer. As shown in Fig. 5, there is an illustration of the shifted window approach. First, a regular window partitioning scheme is adopted in layer Layer (the feature map is divided evenly into 2×2 windows of size 4×4). Then, it can be noticed that the window partitioning scheme has been altered in layer Layer $+ 1$, where the window partitioning is shifted in such a way that the shifted window contains the features of the original neighboring window.

E. Loss Functions in SWCGAN

Our choice is the relativistic average GAN, which differs from the standard GAN that discriminates real images as 1 and fake images as 0. It estimates the probability that the real image is relatively more realistic than the fake image (see Fig. 6). The loss function of the discriminator is defined as follows:

$$\begin{aligned} L_D = &-\mathrm{E}_{x_r \sim p_{\text{data}}(x_r)}\left[\log(1 - D_{\text{RA}}(x_r, z_f))\right] \\ &-\mathrm{E}_{z_f \sim p_z(z_f)}\left[\log(1 - D_{\text{RA}}(z_f, x_r))\right] \end{aligned} \tag{3}$$

where z_f is the high-resolution image by the generator $G(x_i)$, x_i is the input low-resolution image, and x_r is the corresponding real high-resolution image. $x_r \sim p_{\text{data}}$ represents the distribution pattern of the real data, and $z_f \sim p_z(z_f)$ is similar to $x \sim p_{\text{data}}$.

We set the training objective of the generator to minimize the joint loss, which consists of the content loss and the adversarial loss. The loss function of the generator is defined as follows:

$$L_G = L_{\text{cont}} + \lambda L_{\text{adv}} \tag{4}$$

$$L_{\text{cont}} = z_f - x_r{}_1 \tag{5}$$

$$\begin{aligned} L_{\text{adv}} = &-\mathrm{E}_{x_r \sim p_{\text{data}}(x_r)}\left[\log(1 - D_{\text{RA}}(x_r, z_f))\right] \\ &-\mathrm{E}_{z_f \sim p_z(z_f)}\left[\log(D_{\text{RA}}(z_f, x_r))\right] \end{aligned} \tag{6}$$

where the content loss L_{cont} is the L_1 pixel loss to evaluate the mean square error between a generated high-resolution image and real one, the form of the adversarial loss for the generator corresponds to the relativistic average GAN, and λ is a hyperparameter.

III. Results

In this section, we first list the environment of the experiment. In addition, to evaluate the proposed method, we present the results of the comparison between the proposed methods and other methods. Finally, the application results of the proposed methods are shown.

A. Experimental Environment

The details of the experimental environment are listed in Table I.

B. Evaluation of the Proposed Methods

1) Experimental Dataset and Model Training: We selected the commonly used remote sensing dataset "UCMerced" [30] as the experimental dataset for the super-resolution reconstruction task. There are many classes of images in the UCMerced dataset

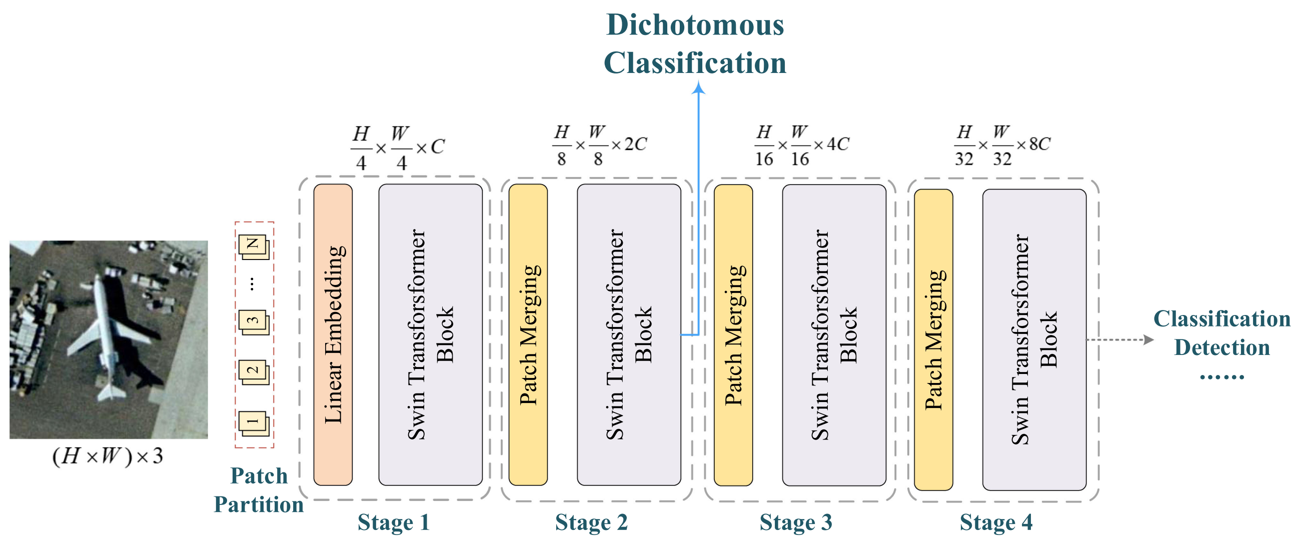


Fig. 4. Illustration of the simplified swin transformer (discriminator).

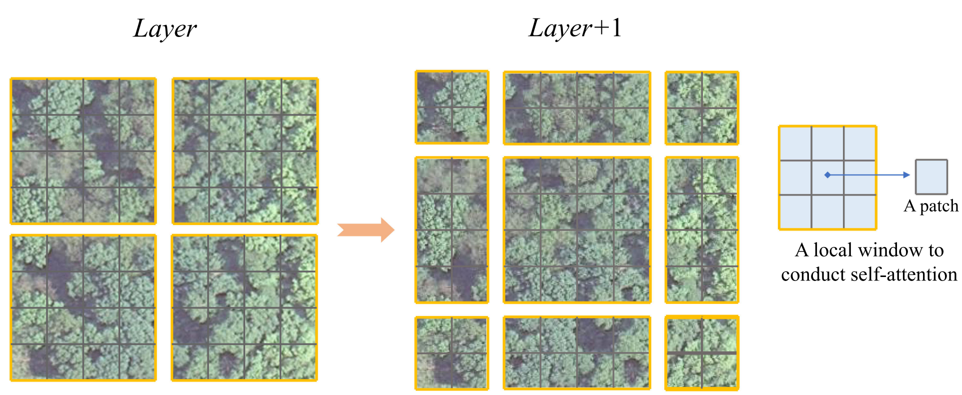


Fig. 5. Illustration of the shifted window approach.

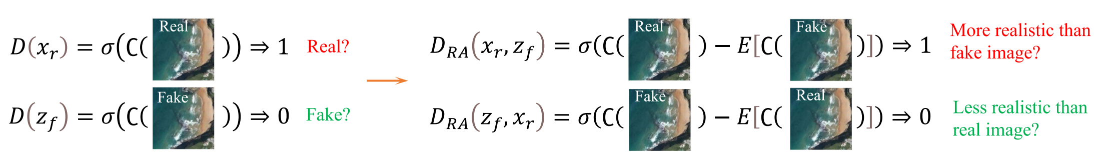


Fig. 6. Illustration of the relativistic average discriminator.

TABLE I
PLATFORMS USED FOR TESTING

Specifications	**Details**
CPU	Intel Xeon 5118
CPU Frequency	2.30 GHz
CPU RAM	128 GB
GPU	Quadro P6000
CUDA Cores	3840
GPU RAM	24 GB
OS	Window 10
Python	Version 3.7
PyTorch	Version 1.8
CUDA version	Version 10.2

with 21 classes (including forest, buildings, beach, and so on), which contain most of the remote sensing scenes. Each class has 100 images, and all images are 256×256 in size. In the training process, we selected 10% of the UCMerced images as the validation set and the remainder as the training set. Furthermore, we use the bicubic algorithm [31], [32] to downsample the UCMerced dataset and obtain the low-resolution dataset as the input.

For the optimizer, the Adam optimizer [33] with the initial learning rate 10^{-4} and the batch size of 8 is used. Specifically, we first pretrain a single generator for 3×10^5 iterations to provide a standard mean square error-based super-resolution model since the training of the generator is based only on the content loss;

TABLE II
RESULTS OF COMPARISON WITH DIFFERENT METHODS ON THE UCMERCED DATASET

Scene Class	**SRGAN**		**EDSR**		**LGCNet**		**RCAN**		**CDGAN**		**EEGAN**		**HSENet**		**SWCGAN (ours)**		**SWCG (ours)**	
	PNSR	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**	**PNSR**	**LPIPS**
Agricultural	25.13	0.479	26.05	0.531	25.94	0.556	26.08	0.519	25.00	0.493	23.27	0.398	23.88	0.429	27.68	0.234	30.23	0.217
Airplane	25.98	0.172	27.39	0.249	26.72	0.299	27.76	0.236	26.42	0.140	30.21	0.204	30.75	0.199	25.17	0.163	26.44	0.240
Baseballdiamond	30.41	0.203	32.88	0.300	32.53	0.308	33.09	0.295	31.02	0.172	34.51	0.261	34.85	0.287	32.30	0.192	33.77	0.302
Beach	32.43	0.198	35.30	0.240	35.07	0.244	35.55	0.234	32.94	0.155	32.01	0.274	32.32	0.297	32.09	0.206	33.29	0.280
Buildings	24.44	0.169	26.10	0.220	25.28	0.277	26.41	0.197	24.66	0.153	24.10	0.251	24.53	0.232	24.76	0.164	26.28	0.181
Chaparral	22.95	0.198	24.81	0.435	24.48	0.430	24.90	0.427	23.69	0.220	24.75	0.377	25.58	0.348	23.24	0.204	24.77	0.403
Denseresidential	24.70	0.179	26.35	0.251	25.67	0.300	26.78	0.228	25.08	0.161	25.07	0.291	25.67	0.275	24.87	0.155	26.18	0.205
Forest	24.95	0.271	26.60	0.537	26.48	0.514	26.65	0.518	25.46	0.160	27.43	0.470	26.72	0.595	25.45	0.282	26.76	0.548
Freeway	26.56	0.185	28.04	0.252	27.32	0.303	28.48	0.232	26.67	0.170	26.22	0.262	26.73	0.282	27.05	0.150	28.53	0.219
Golfcourse	31.31	0.174	34.78	0.269	34.62	0.271	34.98	0.272	32.61	0.120	29.89	0.304	31.91	0.406	29.32	0.237	30.77	0.447
Harbor	20.77	0.199	21.72	0.215	20.88	0.273	22.10	0.180	20.40	0.148	23.62	0.155	22.52	0.112	22.30	0.137	23.32	0.185
Intersection	25.06	0.198	26.53	0.282	26.02	0.327	26.91	0.242	25.18	0.194	25.63	0.226	26.12	0.230	26.27	0.164	27.86	0.214
Mediumresidential	24.22	0.200	25.92	0.305	25.32	0.341	26.32	0.280	24.68	0.189	25.43	0.332	25.97	0.322	22.37	0.207	23.80	0.348
Mobilehomepark	21.68	0.207	22.96	0.268	22.17	0.330	23.43	0.234	21.67	0.208	26.35	0.240	27.19	0.230	21.52	0.197	22.84	0.262
Overpass	24.29	0.212	25.46	0.298	24.61	0.376	24.99	0.264	24.41	0.215	25.86	0.222	26.73	0.225	26.72	0.160	28.73	0.220
Parkinglot	20.74	0.178	21.54	0.246	21.08	0.293	22.08	0.177	20.41	0.162	19.92	0.253	22.10	0.208	21.50	0.156	22.69	0.180
River	25.92	0.269	27.47	0.424	27.39	0.429	27.49	0.412	26.25	0.251	26.39	0.406	26.53	0.473	26.58	0.244	27.91	0.436
Runway	27.91	0.184	29.37	0.247	28.46	0.297	29.42	0.245	28.14	0.158	28.48	0.256	29.29	0.248	28.57	0.158	30.16	0.224
Sparseresidential	26.88	0.185	29.05	0.303	28.58	0.311	29.26	0.300	27.53	0.163	26.93	0.296	27.36	0.297	27.30	0.229	28.73	0.391
Storagetanks	28.42	0.187	30.39	0.241	29.76	0.289	30.46	0.229	28.55	0.162	29.36	0.199	29.37	0.206	27.62	0.194	29.40	0.266
Tenniscourt	27.58	0.183	29.65	0.227	28.99	0.277	30.01	0.214	27.91	0.162	25.79	0.267	26.39	0.266	26.32	0.165	27.83	0.211
Mean	25.82	0.211	27.54	0.302	27.02	0.336	**27.77**	0.283	26.13	**0.193**	26.72	0.283	27.26	0.294	26.14	**0.190**	**27.63**	0.285
Standard deviation	3.142	**0.067**	3.691	0.096	3.795	0.081	3.659	0.100	3.391	0.076	3.145	0.075	3.117	0.106	**3.015**	**0.038**	**3.116**	0.103

A higher PNSR means better result; a higher LPIPS means lower result. It should be noticed that the proposed SWCG is the pretrained generator based on the content loss, which lacks adversarial training with the discriminator compared to the SWCGAN. Red for the first, blue for the second.

on the other hand, it provides better pretraining weights for subsequent adversarial training. Then, based on the relativistic average GAN, the pretraining generator with our discriminator is trained, where the number of iterations is 2×10^5, and the learning rate cuts in half every 7×10^4 iterations. Finally, we obtained the following two versions of the model: 1) the standard mean square error-based super-resolution model (SWCG) and 2) the SWCGAN.

2) Evaluation Metrics: To comprehensively evaluate the performance of the proposed method, we use two evaluation metrics with different focuses. The first metric is the peak signal-to-noise ratio (PSNR), which is the most widely used evaluation metric for images. The PSNR is based on the error between the corresponding pixel points, and minimizing the L_1 loss is equivalent to maximizing the PSNR. The higher PNSR means that the closer the generated image is to the real image, and the better effect of the super-resolution reconstruction. The equation of PSNR is as follows:

$$\mathrm{PNSR} = 10\log_{10}(\mathrm{Max}_I^2/\mathrm{MSE}) \tag{7}$$

where Max_I^2 is used to represent the possible maximum pixel value of the image, and MSE is the mean square error.

However, the PSNR does not consider the visual characteristics of the human eye, resulting in images with high PNSR often being evaluated as low quality [34], [35]. Therefore, the additional evaluation metric, learned perceptual image patch similarity (LPIPS) [36], is selected, which is more consistent with human perception than PNSR. The LPIPS evaluates the perceptual similarity of the image by a deep learning model, where an L_2 distance between the real image x and the corresponding generated super-resolution image x_0 is calculated

$$\mathrm{LPIPS}\left(x, x_0\right) = \sum_l \frac{1}{N_l} \left\| \omega_l \odot \left(\phi(x)_l - \phi\left(x_0\right)_l\right) \right\|_2^2 \tag{8}$$

where $\phi(\cdot)_l$ is a feature space constructed from a well-trained lth layer CNN, and N_l represents the number of elements in $\phi(\cdot)_l$. ω_l is a learned weight vector and $\odot$ is the channel-wise product operation. The lower LPIPS represents the better effect of super-resolution reconstruction.

3) Comparative Results: To further evaluate the performance of the proposed SWCG and SWCGAN, we compared our methods with some advanced super-resolution models, including RCAN [37], SRGAN [16], EDSR [38], LGCNet [39], CDGAN [17], EEGAN [18], and HSENet [40], on the UCMerced dataset.

Table II lists the evaluation results of different algorithms on the UCMerced dataset. We calculated the PNSR, reflecting the mean pixel error, and LPIPS, reflecting the perceptual similarity of super-resolution reconstruction methods on 21 categories of images separately due to the different learning difficulties of different images, and obtained the total average values. For the average PNSR and LPIPS, the proposed SWCGAN receives the best score on the LPIPS metric, and the proposed SWCG receives the second score on the PNSR metric. It is worthwhile to note that these algorithms that minimize pixel loss as a training goal, including RCAN, EDSR, LGCNet, HSENet, and SWCG, can achieve high PNSR, yet their performances on LPIPS are unsatisfactory. In contrast, GANs that incorporate adversarial

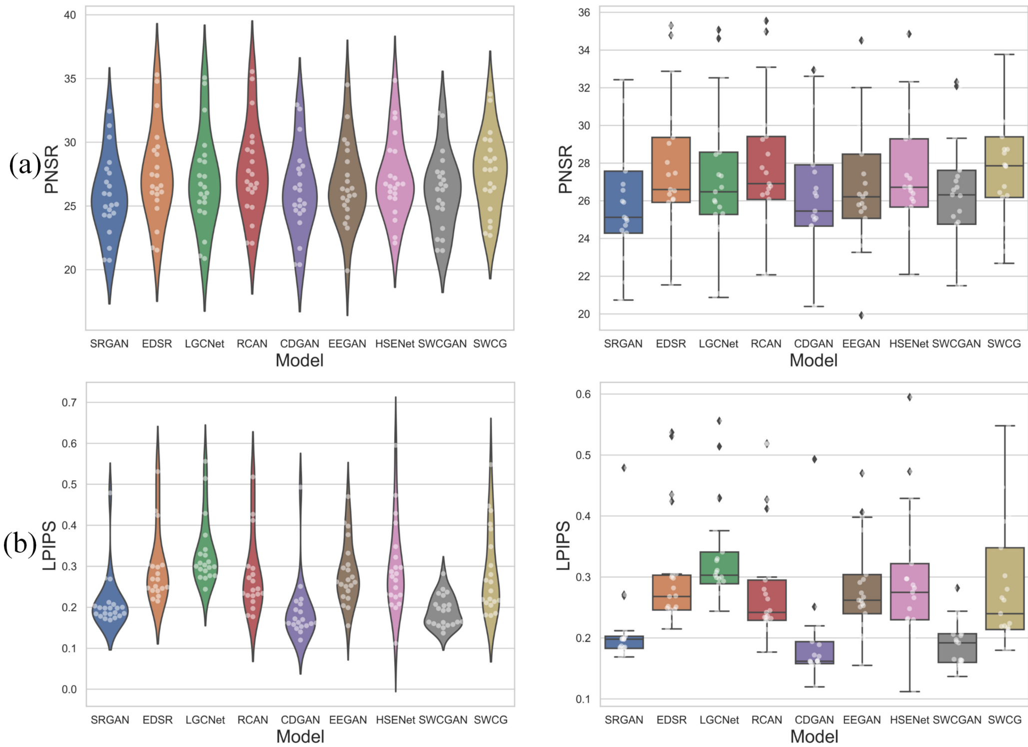


Fig. 7. PNSR and LPIPS distributions of different models on all scenes of the UCMerced dataset. (a) PNSR distribution of different models. (b) LPIPS distribution of different models. (White dot represents the PNSR or LPIPS of the corresponding model on a scene).

loss tend to perform well on the LPIPS metric and obtain lower scores on the PNSR metric. For these GANs, SRGAN obtains the worst performance on the PNSR metric, EEGAN obtains the best performance on the PNSR metric, and the proposed SWCGAN has the best score on the LPIPS metric and the second score on the PNSR metric.

On the other hand, for different scenarios, these deep-learning-based super-resolution algorithms show a significant preference. For example, all algorithms can obtain higher PNSR (>30) on both Baseballdiamond and Beach scenes, and all algorithms obtain lower PNSR (<24) on Harbor scene. Moreover, compared with other algorithms, the proposed SWCGAN obtains the best LPIPS for Denseresidential and Freeway scenes, and the proposed SWCG obtains the highest PNSR for Overpass, Parkinglot, River, and Runway scenes. As shown in Fig. 7, for different scenarios, the values of the proposed SWCGAN are more concentrated around the mean value, whether based on the PNSR metric or the LPIPS metric, with fewer outliers, which means that the performance of the proposed SWCGAN is the most stable compared to other algorithms. As listed in Table II, the proposed SWCGAN has the smallest standard deviation, which also reflects that the SWCGAN has the best stability.

Specifically, as illustrated in Fig. 8, the two examples, Baseballdiamond33 and Golfcourse35, are chosen to compare the details of the different algorithms. In Fig. 8(a), the proposed SWCGAN with the lowest PNSR and the best LPIPS shows the most details and the best perceptual effects. In Fig. 8(b), both SWCGAN and CDGAN with better LPIPS show better perceptual effects. Furthermore, except for SWCGAN and CDGAN, other algorithms exhibit the same image style for the super-resolution reconstruction task, i.e., the borders of objects are not sufficiently clear and the transition between pixels is overly smooth, causing the image to appear blurry.

C. Ablation Studies

According to the comparative results, the proposed SWCGAN obtained the best score on the LPIPS metric, and the proposed SWCG obtained the second score on the PNSR metric. In this section, to further verify the effectiveness of the STB in the proposed RDSTB, the ablation experiments are established by replacing the STB with a convolutional block. Moreover, to demonstrate the impact of the pretraining generator (i.e., SWCG), an experiment is added without pretraining. Finally, to verify the effectiveness of the simplified swin transformer as a discriminator, an experiment using the complete swin transformer as a discriminator is added.

1) ConvGAN: The STBs in the generator of the SWCGAN are replaced by convolution blocks, and the simplified swin transformer is used in the discriminator.

2) ConvG: The STBs in the generator of the SWCG are replaced by convolution blocks, and the discriminator is removed.

3) SWCGAN (N): The proposed SWCGAN without pretraining.

4) SWCGAN (all): the complete swin transformer is used as a discriminator in the SWCGAN.

As listed in Table III, compared with SWCGAN and SWCG, the corresponding models (ConvGAN and ConvG) perform worse on both PNSR and LPIPS metrics when STBs are replaced with convolutional blocks. Compared with the standard

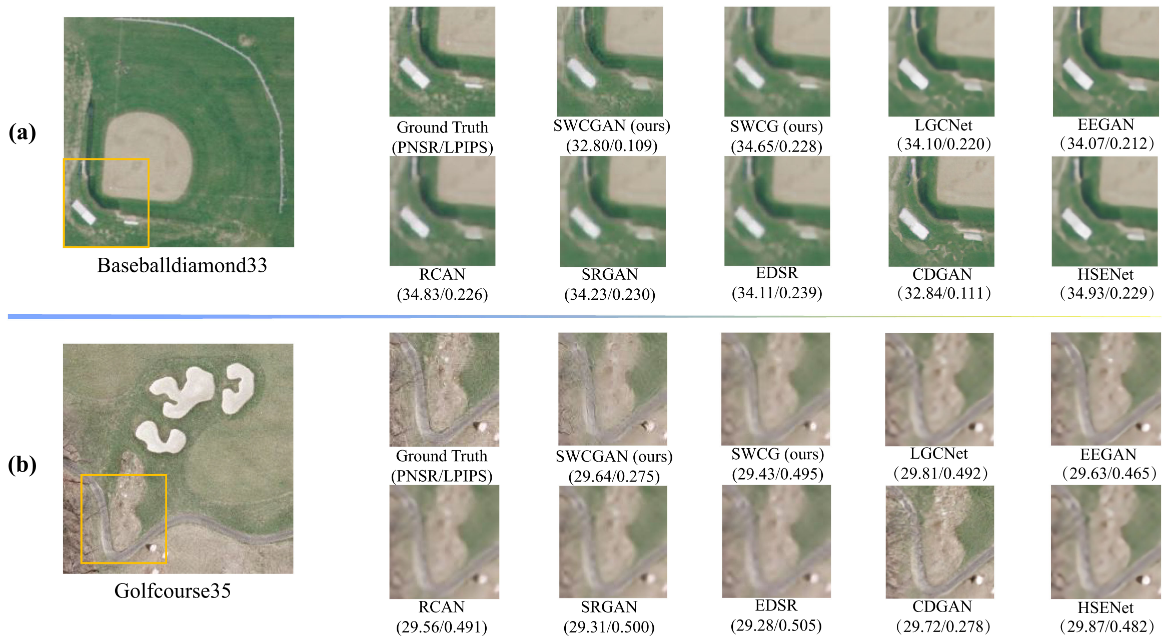


Fig. 8. Detailed comparison of the outputs with different methods. (a) Baseballdiamond. (b) Golfcourse.

TABLE III
RESULTS OF THE BLATION EXPERIMENTS

Model	PNSR	LPIPS
SWCGAN	26.14	0.190
SWCGAN (N)	25.81	0.183
SWCGAN (all)	26.21	0.186
ConvGAN	25.53	0.243
SWCG	27.63	0.285
ConvG	26.57	0.342

All models are evaluated on the UCMerced dataset.

mean square error-based super-resolution model ConvG, due to the effect of adversarial loss, ConvGAN exhibits the same characteristics as other GANs, i.e., lower scores on the PNSR metric and better performance on the human perceptual metric LPIPS.

On the other hand, SWCGAN (N) performs excellently on the LPIPS metric compared to SWCGAN, yet performs poorly (< 26) on the PNSR metric, which is caused by the overpowering influence of the discriminator on the generator. Therefore, a pretraining generator is necessary to make the capabilities of SWCGAN more comprehensive.

Furthermore, SWCGAN (all) with higher complexity only achieves a slightly better performance than SWCGAN. When the input size is 64×64, the floating point operations (FLOPs) of SWCGAN (all) (26.1 G) are 17 % larger than that of SWCGAN (22.3 G), and the FLOPs of the complete swin transformer (5.7 G) are 3 times that of the simplified swin transformer (1.9 G). These experimental results demonstrate the effectiveness of the simplified swin transformer as a discriminator.

D. Application of the Proposed Methods

To verify that the proposed algorithms can be applied to real satellite remote sensing images, we test the super-resolution effect of our proposed SWCGAN and SWCG on a real-world multispectral image from WorldView-4 (WorldView-4 is capable of acquiring satellite images with a panchromatic resolution of 0.3 m and a multispectral resolution of 1.24 m and is used to test the proposed method), where NIQE [41] is chosen as the evaluation metric. Due to the lack of real high-resolution images, we use the nonreference evaluation metric NIQE (a lower score means a better output for super-resolution reconstruction), which is different from PNSR and LPIPS.

As shown in Fig. 9, before super-resolution reconstruction, the NIQE of the original low-resolution image is 17.049. After super-resolution reconstruction, the quality of the image is significantly improved, and the proposed SWCGAN and SWCG obtained the second and first scores compared to other models. When the real satellite remote sensing image (400×400) is used as input, the proposed SWCGAN and SWCG provide more improvement for image quality compared to other models, which means that it is important to capture long-range dependencies between pixels for the real satellite remote sensing image.

IV. DISCUSSION

A. Advantages of the Proposed SWCGAN

The advantage of this method is that it overcomes the inherent drawbacks of the convolutional layer by introducing the swin transformer based on shifted window self-attention, and the proposed SWCGAN achieves the best score in the LPIPS metric and performs better than other GANs in all metrics. For different scenarios, the proposed method has the smallest standard deviation, which means that it has the best stability. Specifically, since the introduction of swin transformer overcomes the shortcomings of convolutional layers, it allows the generator to better learn the relationship between different regions of the image, which leads to the generated images with clear object boundaries, unlike

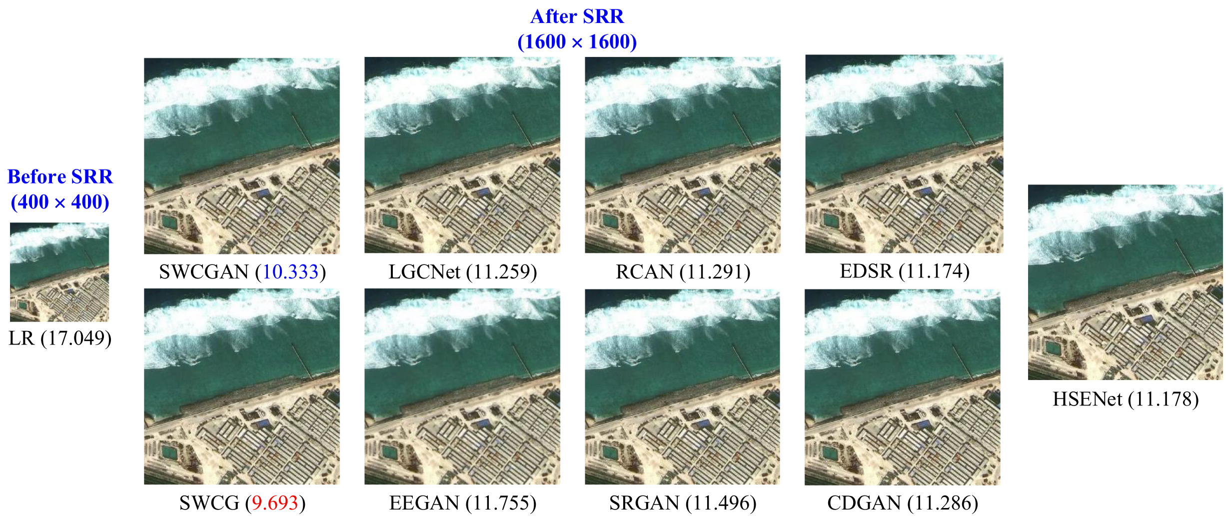


Fig. 9. Application of the proposed methods on a real-world multispectral image for the super-resolution reconstruction. (Red for the first, blue for the second).

TABLE IV
PARAMETERS, FLOPS, AND GPU RUNTIME OF STANDARD MEAN SQUARE ERROR-BASED SUPER-RESOLUTION MODELS

Model	Parameters	FLOPs	GPU Runtime
LGCNet	193k	69.7G	0.096s
EDSR	43M	1130G	0.383s
RCAN	16M	359G	1.211s
HSENet	5.4M	73.3G	1.328s
SWCG (ours)	3.8M	77.6G	0.894s

The input size is 125×125.

other algorithms where the transition between different objects is too smooth, causing the image to appear blurry. Moreover, the swin transformer discriminator is essential for discriminating the generated image from the real image based on fine features, which rights the training objective in such a way that the style of the generated image converges to that of the real image instead of only reducing the error between pixels.

On the other hand, our proposed RDSTB that constitutes the deep feature extractor has an excellent performance to the extent that the proposed SWCG (i.e., the generator of SWCGAN which is trained using only pixel loss) achieves a level close to the state-of-the-art using only four layers of RDSTB. As listed in Table IV, in the standard mean square error-based super-resolution models, the proposed SWCG with significantly lower parameters and FLOPs than those of the RCAN obtains super-resolution performance close to that of the RCAN, and it outperforms the EDSR with exceptionally large parameters and FLOPs.

B. Shortcomings of the Proposed SWCGAN

The shortcoming of the proposed method is that its training time is longer than that of general CNNs due to the introduction of the swin transformer. Several investigations [23] have shown that networks based on self-attention mechanisms require more data and training time than general CNNs for image processing tasks due to the lack of convolutional inductive bias, i.e., the capability to presuppose some necessary assumptions for the problem. In addition, the usage of swin transformer causes a significant increase in complexity. As listed in Table IV, the results show that the SWCG with middle complexity is still much higher than the lightweight network LGCNet even when the lightweight architecture is chosen.

Moreover, although the proposed SWCGAN performs the best in GANs, its performance is still unsatisfactory compared with the standard Mean Square Error-based super-resolution models according to the PNSR metric due to the effect of adversarial training on the loss. Therefore, for super-resolution reconstruction tasks requiring high PNSR, the abovementioned issue can be addressed by adjusting the hyperparameter λ in the loss function of the generator.

C. Outlook and Future Work

In the future, we will continue to use the proposed RDSTB to build a large-scale model to further improve the super-resolution performance of SWCGAN. We believe that the proposed RDSTB can form a deeper and larger network to obtain better super-resolution performance due to the dense connectivity and residual structure. Furthermore, as a feature extraction block of images, the proposed RDSTB will be applied to remote sensing image processing tasks, including classification, recognition, and semantic segmentation. Finally, for the problem of sparse remote sensing image data, we will consider training deep-learning-based models using self-supervised [42] or unsupervised methods [43], [44] in the future.

On the other hand, the superior performance of pure swin-transformer-based models in the field of computational vision has been proven. Thus, we will develop a GAN-based super-resolution network composed of pure swin transformer blocks.

V. CONCLUSION

In this article, we proposed a GAN by combining the advantages of the swin transformer and convolutional layers for super-resolution reconstruction, i.e., SWCGAN. The essential idea behind the proposed method is to generate high-resolution images by a generator network with a hybrid of convolutional and swin transformer layers and then use a pure swin transformer discriminator network for adversarial training. In this proposed SWCGAN,

1) we used a convolutional layer for shallow feature extraction;
2) we proposed the RDSTB to extract deep features of the image for upsampling to generate high-resolution images;
3) we used a simplified swin transformer as the discriminator for adversarial training.

To demonstrate the performance of the proposed methods, we designed an experiment on the UCMerced dataset. The results indicate that the proposed SWCGAN outperforms other state-of-the-art methods in most metrics and performs best on the LPIPS metric compared with other GANs. The ablation experiments suggest the effectiveness of the STB in the proposed RDSTB. Finally, the proposed SWCGAN is applied to a real satellite remote sensing image, and the quality of the remote sensing image is further improved compared to other models.

References

[1] A. Tatem, H. Lewis, P. Atkinson, and M. Nixon, "Super-resolution target identification from remotely sensed images using a hopfield neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 4, pp. 781–796, Apr. 2001.

[2] H. Lin, Z. Shi, and Z. Zou, "Fully convolutional network with task partitioning for inshore ship detection in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1665–1669, Oct. 2017.

[3] L. Gao *et al.*, "Road extraction using a dual attention dilated-linknet based on satellite images and floating vehicle trajectory data," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10428–10438, Sep. 2021.

[4] S. Scepanovic, O. Antropov, P. Laurila, Y. Rauste, V. Ignatenko, and J. Praks, "Wide-area land cover mapping with sentinel-1 imagery using deep learning semantic segmentation models," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 10357–10374, Sep. 2021.

[5] J. Yang, J. Wright, T. S. Huang, and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861–2873, Nov. 2010.

[6] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *J. Opt. Soc. Amer. A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[7] L. Zhang and X. Wu, "An edge-guided image interpolation algorithm via directional filtering and data fusion," *IEEE Trans. Image Process.*, vol. 15, no. 8, pp. 2226–2238, Aug. 2006.

[8] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 2, pp. 295–307, Feb. 2016.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, "Densely connected convolutional networks," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[10] T. Tong, G. Li, X. Liu, and Q. Gao, "Image super-resolution using dense skip connections," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4809–4817.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[12] J. Kim, J. Lee, and K. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1646–1654.

[13] Y. Tai, J. Yang, and X. Liu, "Image super-resolution via deep recursive residual network," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2790–2798.

[14] I. Goodfellow *et al.*, "Generative adversarial nets," *Adv. Neural Inf. Process. Syst.*, 2014, vol. 3, pp. 2672–2680.

[15] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (GAN) based change detection network for optical and SAR remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 179, pp. 14–34, 2021.

[16] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[17] S. Lei, Z. Shi, and Z. Zou, "Coupled adversarial training for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3633–3643, May 2020.

[18] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.

[19] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5999–6009.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 4171–4186.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[22] L. Duong, N. Le, T. Tran, V. Ngo, and P. Nguyen, "Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning," *Expert Syst. Appl.*, vol. 184, 2021, Art. no. 115519.

[23] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–21.

[24] K. Pannerselvam, "Adaptive parking slot occupancy detection using vision transformer and LLIE," in *Proc. IEEE Int. Smart Cities Conf.*, 2021, pp. 1–7.

[25] Z. Liu *et al.*, "Stransfuse: Fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, vol. 14, pp. 10012–10022.

[26] V. Hajihashemi, H. Najafabadi, A. Gharahbagh, H. Leung, M. Yousefan, and J. Tavares, "A novel high-efficiency holography image compression method, based on HEVC, wavelet, and nearest-neighbor interpolation," *Multimedia Tools Appl.*, vol. 80, no. 21–23, pp. 31953–31966, 2021.

[27] M. Kalideen and B. Tugrul, "Outsourcing of secure k-nearest neighbours interpolation method," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 4, pp. 319–323, 2018.

[28] W. Shi *et al.*, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1874–1883.

[29] L. Chen, H. Liu, M. Yang, Y. Qian, Z. Xiao, and X. Zhong, "Remote sensing image super-resolution via residual aggregation and split attentional fusion network," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 9546–9556, Sep. 2021.

[30] Y. Yang and S. Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proc. ACM Int. Symp. Adv. Geographic Inf. Syst.*, 2010, pp. 270–279.

[31] W. Ruangsang and S. Aramvith, "Efficient super-resolution algorithm using overlapping bicubic interpolation," in *Proc. IEEE 6th Glob. Conf. Consum. Electron.*, 2017, pp. 1–2.

[32] Z. Huang and L. Cao, "Bicubic interpolation and extrapolation iteration method for high resolution digital holographic reconstruction," *Opt. Lasers Eng.*, vol. 130, 2020, Art. no. 106090.

[33] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[34] M. Yang and A. Sowmya, "New image quality evaluation metric for underwater video," *IEEE Signal Process. Lett.*, vol. 21, no. 10, pp. 1215–1219, Oct. 2014.

[35] S. Akramullah, *Digital Video Concepts, Methods, and Metrics: Quality, Compression, Performance, and Power Trade-Off Analysis*. Berkeley, CA, USA: Apress, 2014, pp. 101–160.

[36] R. Zhang, P. Isola, A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 6, 2018, pp. 586–595.

[37] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Lecture Notes in Computer Science* (Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics, 11211), Heidelberg, Germany, 2018, 294–310.

[38] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, vol. 7, 2017, pp. 1132–1140.

[39] S. Lei, Z. Shi, and Z. Zou, "Super-resolution for remote sensing images via local-global combined network," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 8, pp. 1243–1247, Aug. 2017.

[40] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5401410.

[41] A. Mittal, R. Soundararajan, and A. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

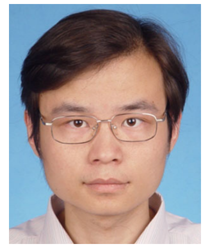
[42] A. Jaiswal, A. R. Babu, M. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, pp. 1–22, 2020.

[43] S. Guan, H. Li, and W.-S. Zheng, "Unsupervised learning for optical flow estimation using pyramid convolution LSTM," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2019, pp. 181–186.

[44] Z. Ma and G. Mei, "Deep learning for geological hazards analysis: Data, models, applications, and opportunities," *Earth-Sci. Rev.*, vol. 223, 2021, Art. no. 103858.

Jingzhi Tu is currently working toward the Ph.D. degree in engineering geology with the China University of Geosciences, Beijing, China.

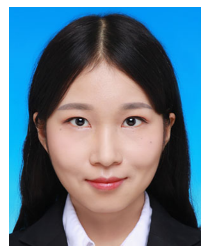
His research interests are in the areas of geohazards prevention, numerical computing, remote sensing, deep learning, and data mining.



Gang Mei received the bachelor's degree in civil engineering and master's degree in geotechnical engineering from the China University of Geosciences, Beijing, China, in 2006 and 2009, and the Ph.D. degree in geological engineering from the University of Freiburg, Freiburg im Breisgau, Germany, in 2014.

He is currently an Associate Professor in Numerical Modeling and Simulation in Civil Engineering with the China University of Geosciences. He has authored or coauthored more than 70 research articles in journals and academic conferences. His main research interests include the areas of numerical simulation and computational modeling, including FEM analysis, GPU computing, data mining, deep learning, and network science and applications.

Dr. Mei has been an Academic Editor for the journals *PeerJ Computer Science*, IEEE ACCESS, and *SN Applied Sciences*, and as a Guest Editor for the journal IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS.



Zhengjing Ma is currently working toward the Ph.D. degree in engineering geology with the China University of Geosciences, Beijing, China.

Her research interests include geohazards prevention, numerical computing, remote sensing, deep learning, and data mining.

Francesco Piccialli received the laurea degree (BSc+MSc) in computer science in 2014, and the Ph.D. in computational and computer sciences from the University of Naples Federico II in 2017. He is currently an Assistant Professor (tenure track) of computer science with the Department of Mathematics and Applications "Renato Caccioppoli," University of Naples Federico II, Naples, Italy. His research interests are focused on data science, machine learning, and Internet of Things (IoT). He is also focusing my research on data mining and data analytics techniques applied on data coming from the IoT world.