Journal of Computers Vol. 34 No. 5, October 2023, pp. 87-101
doi: 10.53106/199115992023103405007

Remote Sensing Image Super-Resolution Using Texture Enhancing Generative Adversarial Network

Shou-Quan Che[1*], Jian-Feng Lu[2]

[1] Liupanshui Normal University, College of Mining and Mechanical Engineering, Liupanshui, 553000, China
Chesq_njtu@163.com

[2] Guizhou University, College of Mechanical Engineering, Guiyang, China
Cme.jflu@gzu.edu.cn

Received 14 September 2022; Revised 25 December 2022; Accepted 13 March 2023

Abstract. Single image super-resolution (SISR) brings excellent improvement in remote sensing applications, which has been widely studied in recent years. A method named TFSRGAN of remote sensing single image super-resolution based on generative adversarial network is proposed in this paper to address the problems of poor reconstruction visual quality and smooth details in traditional algorithms. In the proposed framework, s dense residual connection method is proposed to fuse the deep features from each residual block based on the SRGAN network, and the channel attention mechanism is added into the residual block to combinate the channel information. In addition, the network employs an edge extractor to divide the low-resolution image into low-frequency image and high-frequency image as the input of generator to improve the effect of texture reconstruction. Extensive comparison experiments were performed using AID, UCAS_AOD and China Gaofen-1 datasets, the SR results demonstrate that the proposed TFSRGAN framework outperforms the state-of-the-art algorithms including VDSR, SRGAN and ESRGAN in terms of objective evaluation metrics and subjective visual perception. The ground targets detection experiments represent that the proposed TFSRGAN can significantly improve the effect in remote sensing super-resolution application.

Keywords: remote sensing single image super resolution, generative adversarial network, dense residual connection, channel attention mechanism, texture reconstruction

1 Introduction

High resolution (HR) remote sensing images with abundant detail information are widely used in industry, agriculture, and forestry to geometry map or target detect [1-3]. However, acquiring HR images is not straightforward due to the influence of imaging device performance and environmental conditions. Increasingly, researchers are focusing their efforts on developing the Super-Resolution (SR) technique for remote sensing images. Super Resolution is a term that refers to the process of generating high-resolution images with high quality from low-resolution (LR) images with low quality [4]. Due to the multiple solutions of the pixels in the LR image, the super-resolution methods are ill-posed problems [5]. Currently used SR techniques include Multi-Image Super Resolution (MISR), which utilizes multiple remote sensing images taken from various perspectives to reconstruct a HR result, and Single Image Super Resolution (SISR). Clearly, the MISR method increases the cost of SR significantly and, in most cases, does not have abundant image resources. As a result, the SISR method is more appropriate for general application scenarios, as it acquires an HR image from a single LR image [6-8, 34].

SISR reconstruction attempts to increase LR image resolution without introducing blur or noise. SISR methods are classified into three broad categories: interpolation-based methods [9, 10], reconstruction-based methods [11, 12], and deep learning-based methods. Interpolation-based methods obtain SR results by computing an interpolation formula that considers known pixels and position relationships around a given point. Through multiple frame images and prior information, the reconstructed-based methods establish the reconstruction models between the SR and LR. Both methods mentioned above are conventional and are not discussed in detail in the paper.

Recently, learning-based methods for SR have become more popular and practical, with the goal of learning a model trained on a sample set and then predicting the missing information in LR images. These methods include neighborhood embedding, sparse coding, and deep learning. The inability of neighborhood embedding and

[*] Corresponding Author

sparse coding methods to learn limits their ability to produce desirable results. With the advancement of big data technology, deep learning networks enable SR to overcome numerous constraints and achieve impressive results, establishing it as a research hotspot. Convolutional neural networks (CNNs) have been widely used to address the SR problem due to their strong learning ability. In 2014, Dong et al. [13, 14] applied the CNN model to super-resolution reconstruction image patches and proposed a super-resolution reconstruction convolutional neural network (SRCNN) algorithm that predicted the nonlinear SR pixels value from the LR image patches and outperformed the classic methods significantly. Shi et al. [15] proposed an efficient sub-pixel convolutional neural network (ESPCN) algorithm that utilizes a sub-pixel convolution up-sampling method to improve the SR speed. To accelerate SRCNN, Dong et al. [16] proposed an hourglass-shaped convolutional neural network structure (FSRCNN) with a compact structure and a small convolution kernel.

The deep residual network (ResNet) put forward by He et al. [17] has become a defining feature of CNN. Since then, an increasing number of SISR models have used ResNet to enlarge the network. Kim, et al. [18] pioneered the transition from ResNet to SISR by proposing very deep convolutional networks for image super-resolution (VDSR) to accelerate deep network training. Lim et al. [19] proposed using Enhanced Deep Residual Networks (EDRN) for SISR but eliminate the majority of batch normalization (BN) layers in the original ResNet.

The attention mechanism has been incorporated into image super-resolution reconstruction and yielded impressive results. Zhang et al. [20] put forward super-resolution via the very deep residual channel attention networks (RCAN) algorithm, which derives the weight value from learning the relative importance of various channels and adaptively adjusts the channel characteristics. Recently, SR approaches based on generative adversarial networks (GANs) have been rapidly developed and have achieved a large number of impressive results. In 2017, Ledig et al. [21] proposed a generative adversarial network for image super resolution (SRGAN) as the pioneering work of GAN-based SR methods, in which they used a deep residual network as the generate network (GN) and a classification network as the discriminative network (DN), and proposed a perceptual loss function that contained an adversarial loss and a content loss as the network training loss function. Wang et al. introduced Enhanced super-resolution generative adversarial networks (ESRGAN) [22] as an improvement to SRGAN, which at the time achieved state-of-the-art performance.

In the field of Remote Sensing Single Image Super Resolution (RSISR), increasing CNN- [23-25] and GAN-based methods [26-28] that draw on the network of the above learning-based algorithms have been proposed. On the other hand, in comparison to natural image SR, remote sensing image SR applications frequently involve various additional post-processing steps such as the classification, detection, and measurement of ground targets. However, the target objects typically take up only a few pixels in the remote sensing image. It is critical that the SR results contain abundant texture features that can enhance the target objects attribute description ability. Unfortunately, the remote sensing SR results of existing methods are smooth and lack of texture details. To address the above drawbacks, we investigate a Texture enhancing Generative Adversarial Network (TESRGAN) for remote sensing super-resolution. The following are this paper's primary contributions:

(1) We propose TESRGAN, a GAN-based network use for remote sensing SR reconstruction, which introduces a technique that divides the LR image into low and high-frequency images for use as the parallel inputs of the GN module in the network. Different from the traditional GAN-based methods, the proposed network naturally pays more attention to the texture features and reconstruct the result with abundant and clear details;

(2) We build residual blocks (RBs) with a channel attention (CA) mechanism to combine the network's channel information, which can effectively and adaptively select useful information and emphasize these features in the reconstruction. The CA mechanism improve the SR quality without increasing network depth;

(3) Under the SRGAN baseline, we put forward a dense connection method for residual blocks to combine the layer features in the network in order to take advantage of all the hierarchical features from the LR inputs and effectively prevent gradient vanishing problem;

(4) The proposed TESRGAN method was compared with the classical algorithms and the extensive experiments were performed using the AID, UCSA-AOD datasets and real scene data of China Ggaofen-1. The results demonstrate the competitive performance of proposed network when super-resolving remotely sensed images.

The remainder of this paper is organized as follows. In section 2, we introduce the related GAN and SRGAN work. Then, we describe the proposed TESRGAN method in detail in Section 3. Section 4 dedicates to the experiment details and the comparison of the results with those of other state-of-the-art methods. We present the conclusion and discussion in section 5.

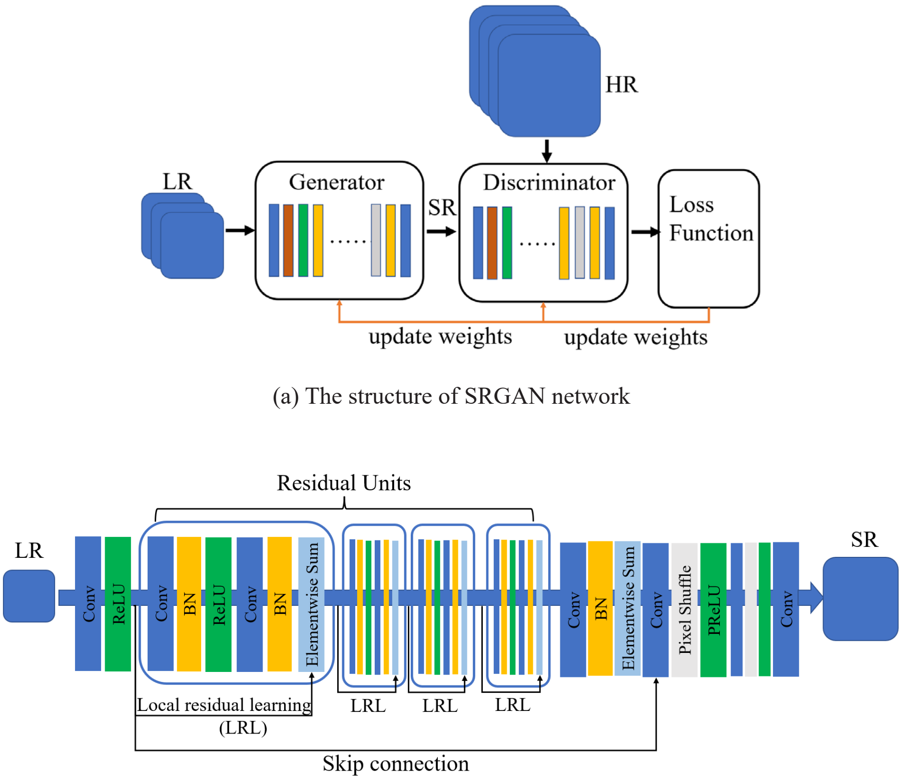
2 Related Work

GAN is a robust deep learning model that is currently under investigation. It was proposed by Goodfellow et al. [29] in 2014 and the model optimizes generation results through a confrontation process between the generative and discriminant modules based on the theory of two-person zero sum games. SRGAN is a kind of GAN-based super-resolution network. SRGAN describes a very deep ResNet architecture that utilizes the concept of GANs to construct a SISR model and produces more photorealistic results than the reconstructions of CNN-based methods [21]. As represented in Fig. 1, the framework of SRGAN contains two main modules, GN and DN. SRGAN uses a deep residual network (SRResNet) with skip connections as the GN to generates the estimation information of pseudo high-resolution images, and uses GN judges the error between the estimation images and the real high-resolution images. The continuous process of cyclic generation and verification iteratively optimize the parameters of the model and the SR results are generated by the GN after training. The antagonism formula of GN and DN is shown in (1).

$$\min{}_G \max{}_D V(D,G) = E_{x\sim P_{data}(x)}\left[logD(x)\right] + E_{z\sim P_z(z)}\left[log\left(1-D\left(G(z)\right)\right)\right]. \tag{1}$$

where is the real HR data and is the input LR data, $D(\cdot)$ represents the probability that the input of the function is judged to be a real data and the $G(\cdot)$ presents the output of the generator.

Additionally, SRGAN defines a novel perceptual loss consisting of an adversarial and a content loss. The content loss is calculated as the Euclidean distance between the feature maps of the generator reconstructions and the VGG19 ground-truth images. The discriminator provides the adversarial loss. ESRGAN used a novel generator to eliminate all BN layers and a relativistic discriminator in place of the standard discriminator [22] to produce more natural-looking reconstructions in natural image sets than SRGAN.



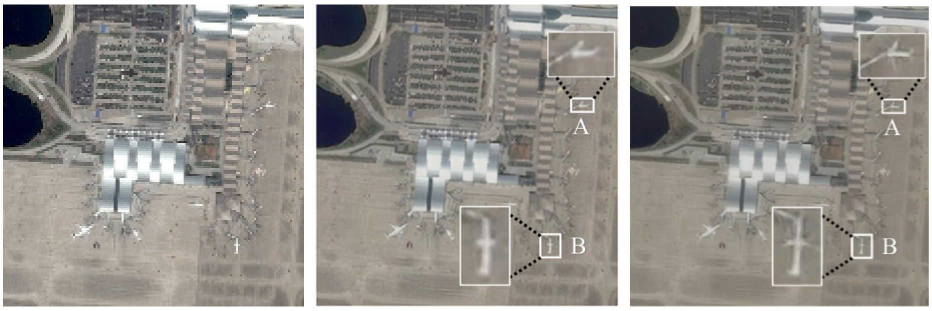
(a) The structure of SRGAN network

(b) The structure of GN in SRGAN

Fig. 1. The architecture of SRGAN and the Generator Network (GN) in SRGAN

3 Proposed Algorithm

The purpose of super-resolution remote sensing is not only to obtain HR images from LR images with respect to color distribution or structural features, but also to facilitate subsequent research on target images include target detection and classification that have always been hot topics in remote sensing applications. As represented in Fig. 2, since the detected object in LR image occupies less than 10*10 pixels, reconstruction result using classical GANs is insufficient to recover the texture features. The result produced by the proposed algorithm in this paper is more visually natural and easier to correctly classified in the detection network.



(a) LR (b) SRGAN (c) Proposed algorithm

Fig. 2. SR performance of SRGAN and the proposed algorithm

[(a) is the LR image, (b) is the SR of SRGAN, and (c) presents the SR of the proposed algorithm. Inputting (b) and (c) SR images into the object detection network YOLO-V3, the airplane patches in the position A and B are scored 0.48, 0.39 and 0.66, 0.57 as the class of the airplane.]

3.1 TESRGAN

TESRGAN, like the majority of GAN-based reconstruction methods, consists of a generator G and a discriminator D. G takes the LR image as input and generates a SR image, then the SR image and ground-truth HR image are synchronously fed into G to judge the authenticity of the SR. To update the weights of the generator and discriminator networks, the loss function is defined and backpropagated. As illustrated in Fig. 3, the input LR image is divided into low frequency content information and high frequency texture information, in contrast to conventional GAN-based methods. The high frequency texture information is yielded by the convolution operation of an edge detection detector and the LR image, and the low frequency information is derived from the LR image by subtracting the high frequency information.

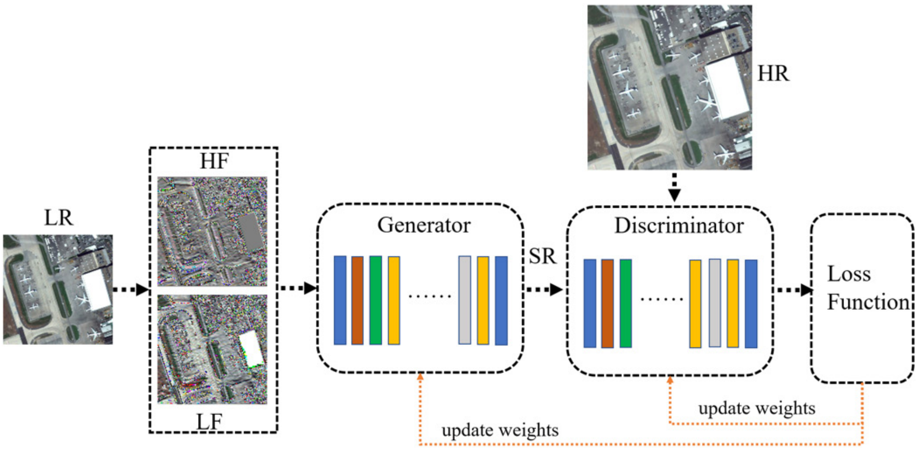


Fig. 3. The general structure of TESRGAN

[HF (high-frequency image) and LF (low-frequency image) were obtained by the convolution operation.]

3.2 Generator Architecture

As above, SRGAN uses a deep residual network (SRResNet) with skip connections as the generator to extract the features and estimate the corresponding HR counterpart for a given LR input image. Indeed, as a feed-forward CNN, the generator omits the connection step and simply integrate the local input and output information in residual blocks, which does not make commendable use of the features extracted from each layer in the deep network. To address the aforementioned issues, a dense residual network (DResNet) with channel attention is proposed in the GN module of TESRGAN. As illustrated in Fig. 4, the generator is composed of three components: preliminary learning, which extracts shallow features, dense residual blocks, which learns deep features and merges all the maps produced by the previous module to improve learning ability without increasing the number of layers, and image reconstruction, which performs image upscaling. Notably, the two input images share the GN parameters.

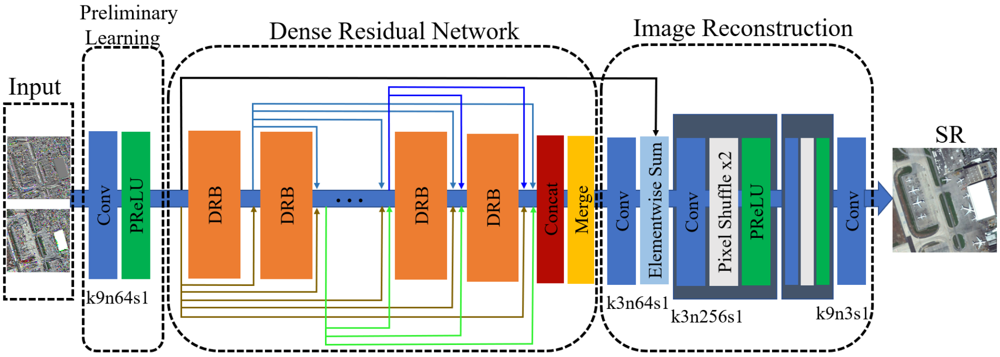


Fig. 4. The generator architecture

The preliminary learning module contains a convolutional layer composed of 64 filters with kernel size 9*9 and stride 1, as well as a nonlinear operator PReLU for extracting the fundamental features. The calculation results of preliminary learning are inputted to DResNet. DResNet is composed of multiple fundamental Dense Residual block (DRB) components, as well as a feature merging and learning process.

DRB is a CNN-based feature extractor whose basic structure is inspired by SRGAN's convolutional layers, except that the BN layer is removed and a channel attention mechanism (CA) is added, as illustrated in Fig. 5. After twice convolution and nonlinear activation processing, the input map is imported into the CA module. The final result obtained by connecting the input map and the output of the CA module via a skip elementwise sum connection.

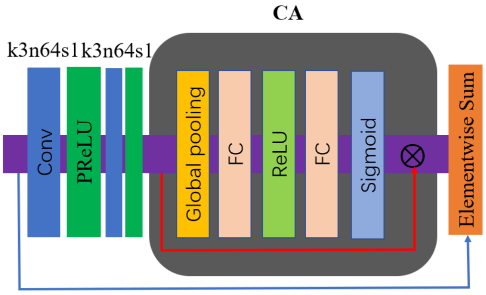


Fig. 5. The structure of DRB (The Operator denotes an element-wise product.)

According to relevant theories and experiments [19, 22, 30], when the network is deep and trained using the GAN framework, BN layers tend to introduce undesirable artifacts and limit generalization ability. In order to modeling the interdependencies of the convolutional features in each channel, the CA mechanism is proposed to enhance the network's presentational power. The Squeeze-and-Excitation block, a new architectural unit introduced in CA, enables the network to perform feature recalibration, selectively emphasizing informative features and suppressing less useful ones [31].

In the CA module, the output is mapped by reweighting the input features in various channels and pays attention to high frequency information. For the input features' size of $H * W * C$, global pooling layer adopts a global average pooling algorithm to calculate a single average value of each channel, determined as follows:

$$g_z = \frac{1}{H * W} \sum_{i=1}^{H} \sum_{j=1}^{W} p_z(i, j). \tag{2}$$

where g_z presents the output and $p_z(i, j)$ denotes a certain value in the position (i, j) of the z-th feature channel.

A $1*1*C$ channel descriptor obtained after the global pooling layer is used as the input of the full connection (FC) layer. The FC layer scales the output with a reduction ratio of r. After nonlinear operating by the function ReLU, another FC layer rescales the output to $1*1*\frac{c}{r}$. The sigmoid function is applied as a gating mechanism to obtain the weight coefficient of each channel that manifests the channel dependence. The process is expressed as follows:

$$w_z = f(W_s \sigma(W_f g)). \tag{3}$$

$f(\cdot)$ denotes the sigmoid function while $\sigma(\cdot)$ denotes the ReLU function. W_s and W_f present the respective weights sets of two FC layers. Finally, the feature of the -th channel is mapped as follows:

$$\hat{g}_z = g_z * w_z. \tag{4}$$

Regarding the DResNet structure, for d-th RB, the input is the result of a convolution after concatenating the $(d - 1)$-th input and output with a concatenation function. This function aims to concatenate all input maps and the convolution with a kernel of 1*1 and a stride of 1, intending to fuse each channel information in each map. We denote the result of preliminary learning module as R_{-1}. For d-th RB, the output is expressed as follows:

$$R_d = F_{RB}(F_{merge}(F_{concat}[R_{d-1}, ..., R_0, R_{-1}])), \tag{5}$$

$$R_0 = F_{RB}(R_{-1}). \tag{6}$$

F_{concat} denotes the concatenation function, F_{merge} signifies the convolution and F_{RB} refers to the calculation process in the RB module.

The same combination and merge process are followed by total K RB modules, and a 3*3 convolution layer further extracts the fusing features. Eventually, a global skip connection is employed to improve network representation ability. The feature learning process result R_F can be expressed as follows:

$$R_F = R_{-1} + F_{conv}(F_{merge}(F_{concat}[R_K, ..., R_0, R_{-1}])). \tag{7}$$

$F_{conv}(\cdot)$ denotes the convolution operation.

We apply an efficient sub-pixel shuffle layer in conjunction with a convolution layer and an activation layer in the image reconstruction module to reconstruct and upscale the input features inspired by ESPCN [15] and SRGAN. This module's feature upscale process is as follows:

$$I_{SF} = w_{pf} * \sigma(F_{conv}(R_F)). \tag{8}$$

Where * denotes the sub-pixel shuffle convolution process and w_{pf} presents the weights of the convolution layer.

The SR image is obtained by convoluting the output of the shuffle layer with a kernel of 9*9 and the output channel of 3 convolution layers.

3.3 Discriminator Architecture

TESRGAN follows the structure of DN in SRGAN. The DN contains eight convolution layers with a 3*3 filter kernel, and the number of filter elements double increase from 64 to 512. Additionally, when the number of features is doubled, convolution layers with a stride of 2 are used to reduce the image resolution. To obtain the classification probability, the resulting feature maps are input into two dense layers and a final sigmoid activation. The structure of DN is depicted in Fig. 6.

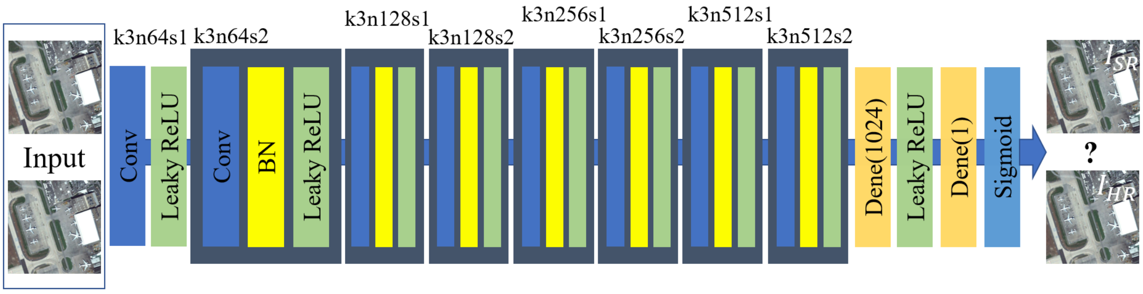


Fig. 6. The DN structure of TESRGAN

3.4 Loss Function

A perceptual loss function that different from those commonly modeled based on MSE is proposed in SRGAN, which has been proved to reconstruct more abundant and natural details in SR. In the proposed algorithm, the loss function is still used. The perceptual loss function is formulated as the weighted sum of a content loss $l^{SR}_{VGG/i,j}$ and an adversarial loss l^{SR}_{Gen}, specified as follows:

$$l^{SR} = l^{SR}_{VGG/i,j} + 10^{-3} l^{SR}_{Gen}. \tag{9}$$

Instead of an assessment of pixel-wise MSE loss that over-smooths the textures, defining the content loss as the feature representational Euclidean distance between reconstruction $G_{\theta G}(I^{LR})$ and the HR I^{HR} in a pre-trained VGG19 network, which represented as follows:

$$l^{SR}_{VGG/i,j} = \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\phi_{i,j}(I^{HR})_{x,y} - \phi_{i,j}(G_{\theta G}(I^{LR}))_{x,y})^2. \tag{10}$$

Where $W_{i,j}$ and $H_{i,j}$ denote the sizes of the feature map within the VGG network respectively. $\phi_{i,j}$ indicates the feature map obtained by the j-th convolution after activation and before the i-th max-pooling layer in the VGG network.

Adversarial loss is a technique proposed in SRGAN for optimizing the network by attempting to fool the DN into favoring solutions that reside on the manifold of HR images. It is defined in terms of the DN's probabilities across all training samples as follows:

$$l_{Gen}^{SR} = \sum_{n=1}^{N} -\log D_{\theta G}(G_{\theta G}(I^{LR})). \quad \textbf{(11)}$$

Where $D_{\theta G}(G_{\theta G}(I^{LR}))$ is the probability of the SR $G_{\theta G}(I^{LR})$ that is an estimation HR image output by DN.

4 Experiment

4.1 Data

The AID dataset, a standard remote sensing scene classification dataset, contains over 10,000 images and more than 30 categories, which are divided into various data subsets for network training and testing in our work. Another high-resolution image dataset, UCSA-AOD, which is widely used in remote sensing target detection, is used to evaluate the proposed method's applicability and superiority. Additionally, real remote sensing data of China Gao-fen1are used in the experiment to avoid model deviation in practice due to the difference between the manual down-sampled image and the real LR image. Following SRGAN, the LR images used for training are generated using bicubic interpolation with a down-sampling factor from the ground-truth images counterparts selected from the datasets.

4.2 Evaluation Metrics

Following SRGAN, we assessed the perceptual quality using a mean opinion score (MOS) test. We assigned 12 volunteers to an independent evaluation task in which they were asked to rate the SR images on a scale of 1 (poor quality) to 5 (excellent quality).

In our work, the SSIM metric is used to represent the structural similarity between the reconstructions and ground truth. Its calculation is shown in Equation (12).

$$SSIM(X,Y) = \frac{(2u_X u_Y + C_1)(2\sigma_{XY} + C_2)}{(u_X^2 + u_Y^2 + C_1)(\delta_X^2 + \delta_Y^2 + C_2)}. \quad \textbf{(12)}$$

Where u_X and u_Y represent the means of the gray values of image X and Y, and δ_X^2 , δ^2 and σ_{XY} represent the variance of the gray values of image X and Y respectively. Generally, $C_1 = (K_1 * L)^2$, $C_2 = (K_2 * L)^2$, $K_1 = 0.01$, $K_1 = 0.03$ and L is the maximum image value. The larger SSIM value, the less image distortion there is. The range of SSIM is from 0 to 1.

Additionally, recent researches indicate that conventional quantitative measures such as PSNR and SSIM are insufficient to assess image quality in terms of the human visual system [21, 32]. A human perception-based measures LIPIS is used in our work. LIPIS is a perceptual distance metric based on deep network feature learning, which is proposed to assess the similarity of image patches and the result being considered closer to human visual measurement, and the details of LIPIS are presented in paper [33]. A higher LIPIS value indicates a greater degree of difference between the SR images and ground truth, while a lower value indicates a greater degree of visual similarity.

4.3 Training Details

We implemented the experiment on the TensorFlow deep learning framework based on the Ubuntu 18.04 operation system. The hardware configuration involved an Intel Core I7-10700K CPU and a GeForce RTX 3080 GPU. We obtained the LR images to be used for training by down-sampling the HR images in the above datasets using bicubic kernel with a down-sampling factor $\gamma = 4$. A training image batch contains 16 random 40*40 cropping sub-images from distinct ignoring classification training images. We used Adam algorithm for optimization. The GN with 15 DRBs were pre-trained based on the MSE loss function with a learning rate of 10^{-4} and 10^{6} update iterations. Then, the pre-trained GN was used as the generator initializer of TESRGAN. The proposed TESRGAN trained with 10^5 update iterations at a learning rate of 10^{-4} and another 10^5 iterations at a lower rate of 10^{-5}.

4.4 Comparative Experiment

We compared TESRGAN to bicubic interpolation method and three other state-of-the-art methods: VDSR [19], SRGAN [21], and ESRGAN [22]. The experiments were conducted on the AID and UCAS-AOD datasets test image tokens. Fig. 7 illustrates the remote sensing LR images SR reconstruction examples from the AID dataset using various methods at a up-scale factor of 4. Compared with other methods, the SSIM, LIPIS and MOS evaluation values of the proposed method TESRGAN are higher, lower and higher respectively, that means the reconstruction results performed better in structure similarity and human visual perception. In other words, the results indicate that the SR results were closer to the HR image. As shown in the Fig. 7 the texture details of the TESRGAN reconstructions were more distinct and abundant, the subjective perceptual quality is improved.

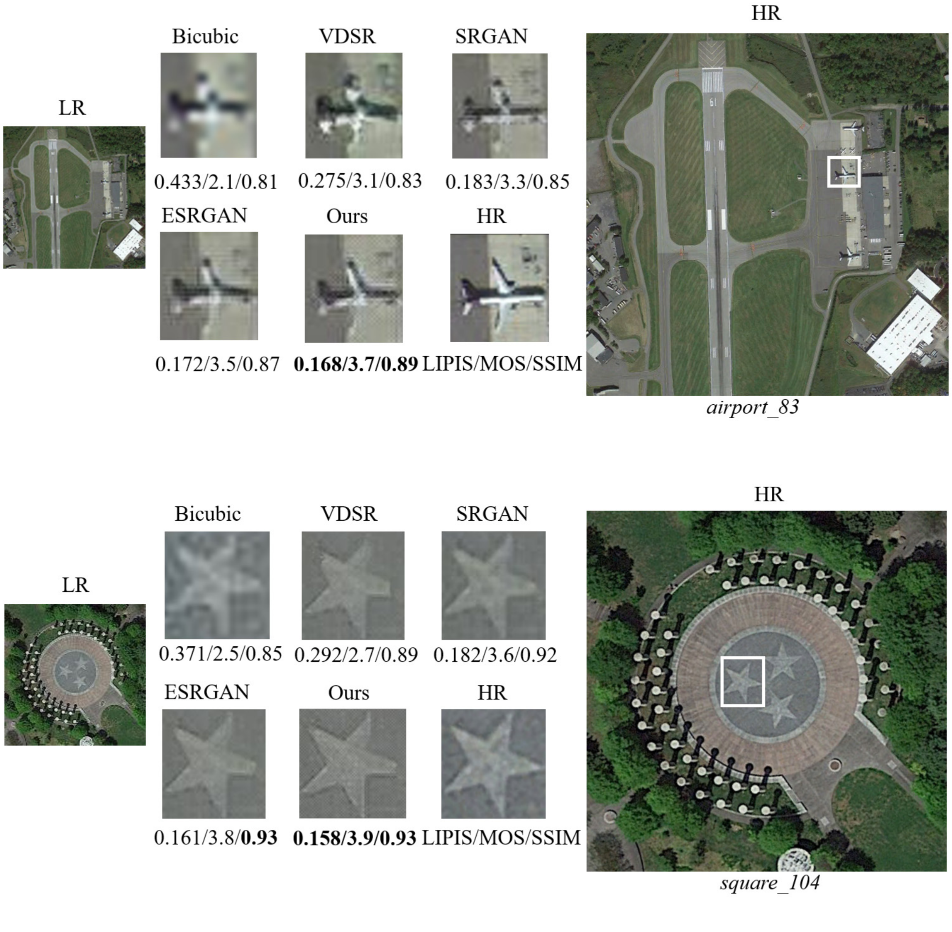


Fig. 7. Reconstruction results comparison of x4 SR in AID dataset (The best result in bold.)

To verify the performance of the proposed algorithm in the dataset, Table 1 summarizes the mean values of LIPIS, MOS and SSIM for the 320 test images included in the AID's sub-datasets, which include airport, bridge, church, farmland, industrial, park, square, and railway station.

Another test approach was evaluated using images in the UCAS-AOD dataset. As illustrated in Fig. 8, the SR results obtained by TESRGAN for target objects on the ground were more detailed and clearer, resulting in higher evaluation scores. Additionally, 100 images tokens from sub-datasets of the UCAS-AOD dataset, including CAR and Neg, were used to score the reconstruction effect of various methods in the evaluation metrics, as shown in Table 2.

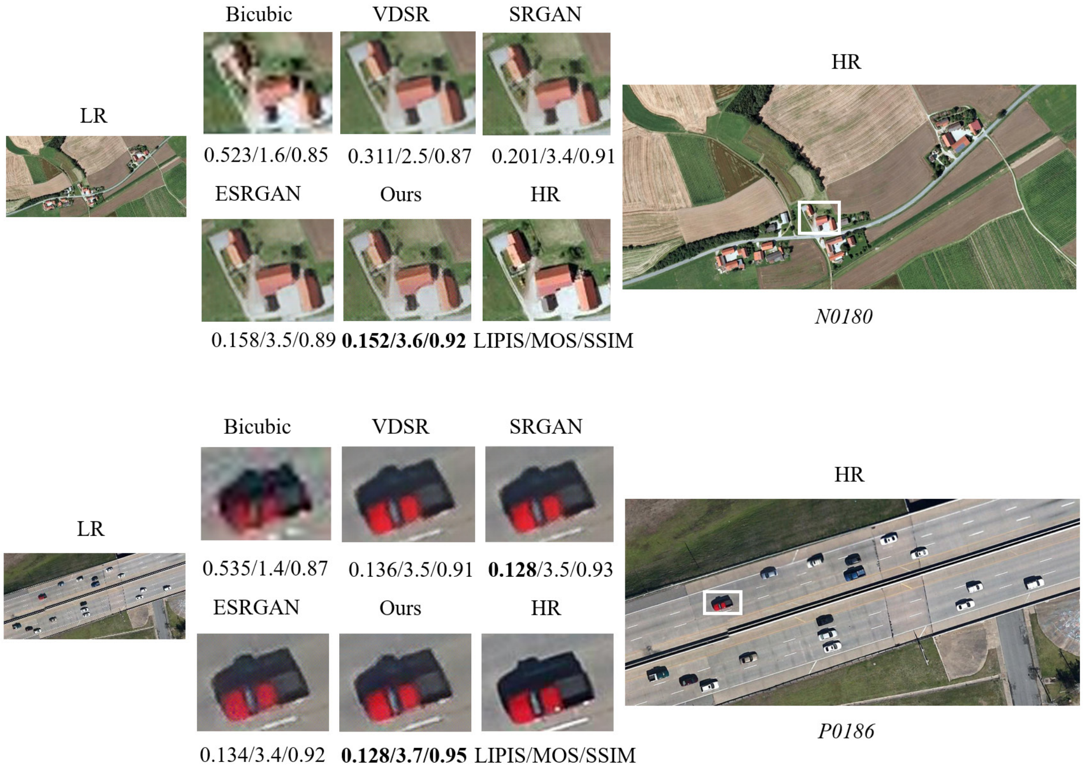


Fig. 8. Reconstruction results comparison of x4 SR in UCAS-AOD dataset (The best result in bold.)

Table 1. Comparison of SR used different methods with x4 upscale in AID dataset

Sub-dataset/Number of images	Bicubic LIPIS/MOS/ SSIM	VDSR LIPIS/MOS/ SSIM	SRGAN LIPIS/MOS/ SSIM	ESRGAN LIPIS/MOS/ SSIM	Ours LIPIS/MOS/ SSIM
Airport/40	0.437/1.5/0.85	0.175/2.4/0.87	0.157/3.1/0.89	0.152/3.4/0.89	**0.151/3.7/0.91**
Bridge/40	0.419/1.6/0.85	0.179/2.6/0.87	**0.156**/3.4/0.89	0.157/**3.6/0.91**	**0.156/3.6/0.92**
Church/40	0.523/1.4/0.83	0.187/2.4/0.84	0.156/3.5/0.88	**0.152**/3.7/0.87	**0.152/3.8/0.89**
Farmland/40	0.412/1.5/0.86	0.151/2.7/0.89	0.155/3.7/0.89	**0.145/3.8/0.91**	0.147/**3.8/0.91**
Industrial/40	0.334/1.4/0.85	0.189/2.8/0.86	0.152/3.6/0.86	0.147/**3.8/0.89**	**0.138/3.8/0.89**
Park/40	0.336/1.5/0.87	0.173/2.7/0.89	0.143/3.6/0.91	**0.132/3.7**/0.91	0.135/**3.7/0.92**
Square/40	0.287/1.5/0.85	0.156/2.8/0.86	0.136/3.6/0.87	0.131/3.5/0.87	**0.129/3.7/0.89**
RailwayStation/40	0.383/1.4/0.86	10.232/2.6/0.87	0.139/**3.7**/0.87	0.136/3.6/0.89	**0.132/3.7/0.90**

Table 2. Comparison of SR used different methods with x4 upscale in UCAS-AOD dataset

Sub-dataset/Number of images	Bicubic LIPIS/MOS/ SSIM	VDSR LIPIS/MOS/ SSIM	SRGAN LIPIS/MOS/ SSIM	ESRGAN LIPIS/MOS/ SSIM	Ours LIPIS/MOS/ SSIM
Car/50	0.215/1.5/0.87	0.163/3.3/0.87	0.157/3.7/0.89	0.155/3.7/0.89	**0.152/3.8/0.90**
Neg/50	0.231/1.4/0.85	10.158/2.6/0.87	0.132/**3.7**/0.87	**0.129**/3.6/0.87	0.131/**3.7/0.89**

It can be observed from Table 1 and Table 2, in SR reconstruction test of the two datasets, the reconstruction results of the proposed method have higher average SSIM value than that of other methods, which means a better structural similarity. Meanwhile, the lower average LIPIS values and higher average MOS values mean the results of the proposed method have a better performance in visual perception, whether in terms of objective indicators or subjective experience.

To validate the proposed algorithm’s performance in the real-world scenario, remote sensing images from China Gaofen-1 satellite were used in the experiment. The Fig. 9 shows the partial SR test results with an up-scaler of 4, and the input LR images and ground truth HR images are taken from the same satellite data but with different resolution. Obviously, the SR results of VDSR lack natural and clear details, SRGAN produces images with improved brightness and contrast, but with some unappealing artifacts. The SR results of ESRGAN are closer to HR images but smooth in details. TESRGAN has better performance in both evaluation metrics and visual perception.

As mentioned before, SR remote sensing images are usually used in the post-processing applications such as target detection and classification. In our work, we train a remote sensing images ground target classification framework base on YOLO-V3 algorithm firstly, then the SR results of each algorithm are input into the network to validate the reconstruction performance of the proposed TESRGAN. Note that we only select part of the images in the AID, UCAS-AOD and Gaofen-1 datasets then cut them randomly to 280*280, and only detect three targets including airplane, car and railway-station. The targets detection of SR images example is represented as Fig. 10. Table 3 presents the detection accuracy of different SR results, 120 images in total. Obviously, the proposed method improves the target detection performance.

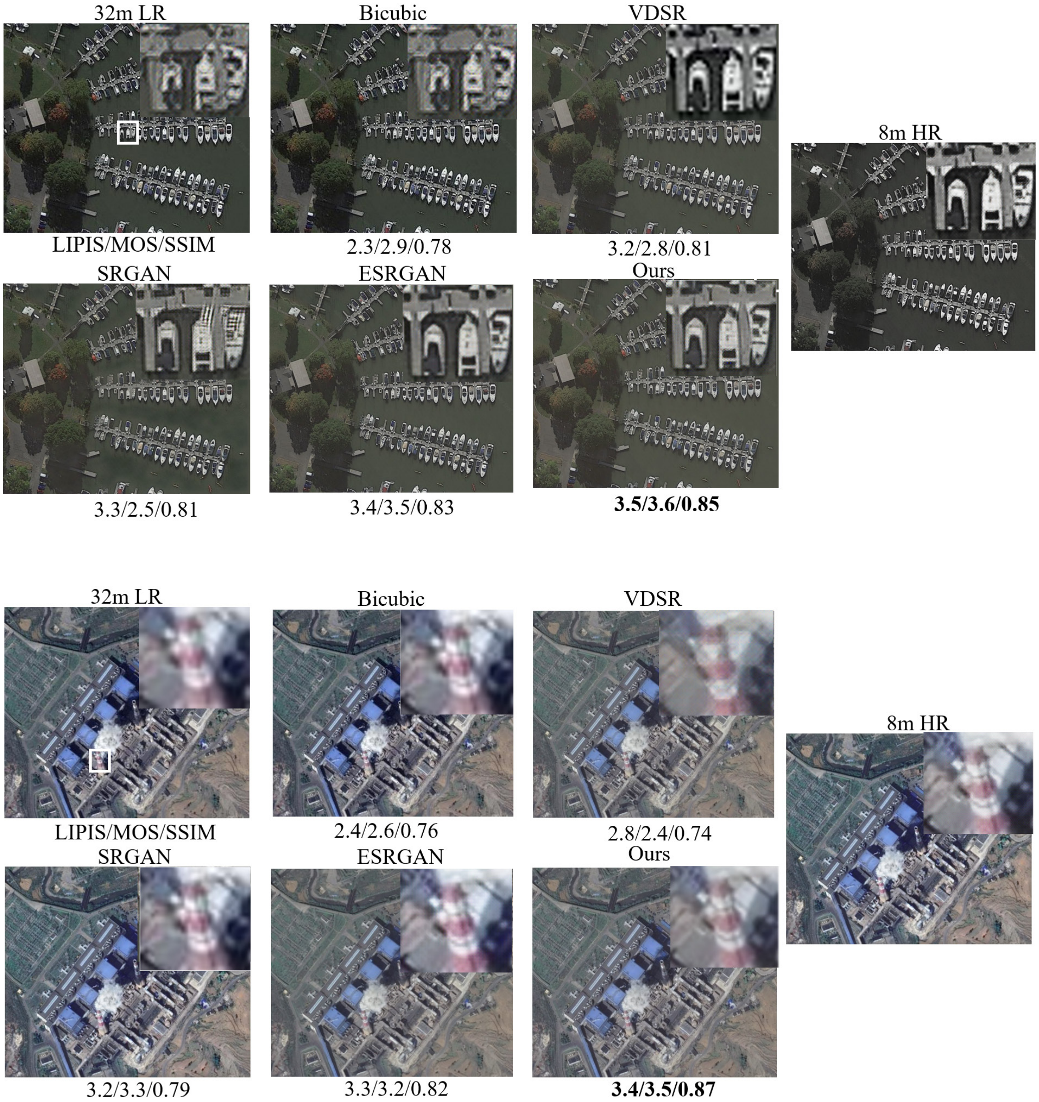


Fig. 9. SR results of real scene data taken from China Gaofen-1dataset

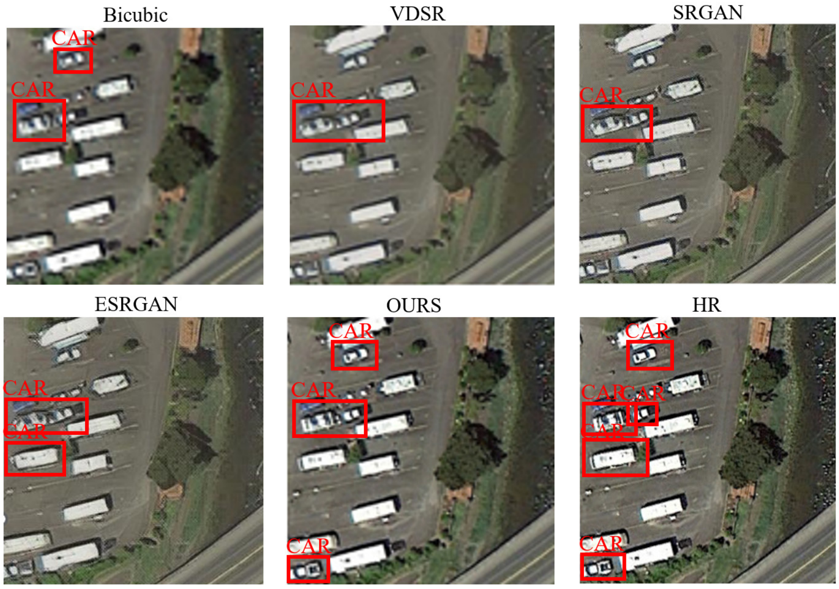


Fig. 10. Targets detection comparison of the SR results reconstructed by different methods

Table 3. Comparison of targets detection of the SR that reconstructed by different methods with x4 upscale

Target/Number of images	Bicubic	VDSR	SRGAN	ESRGAN	Ours	HR
	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)	Accuracy (%)
Car/40	34	52	65	67	69	88
Airplane/40	37	48	58	69	70	92
Railway station/40	58	67	75	78	82	97

4.5 Ablation Study

The DRB modules used in TESRGAN are designed to extract deep features, the network's depth is primarily determined by the number of DRB modules. As illustrated in Fig. 11, the deeper network reconstruction is closer to ground truth, and the LIPIS and SSIM score gradually change slowly after 15 blocks, while training time increase significantly. So, we chose a network with 15 DRB modules after conducting the investigation of training costs and performance on evaluation metrics.

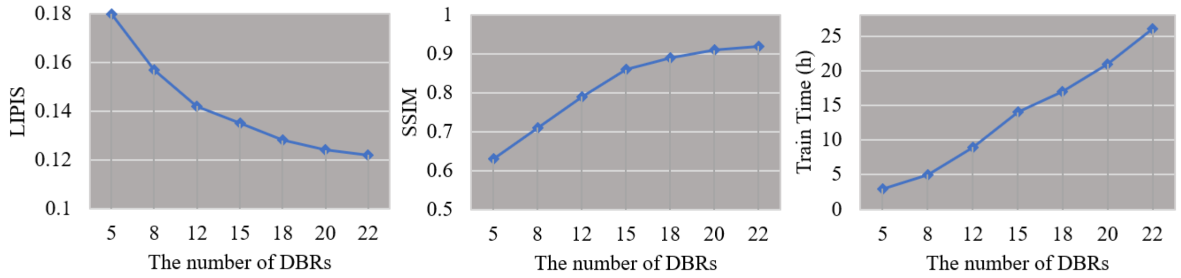


Fig. 11. Dependence of network performance (LIPIS, SSIM and training time) on the number of DRBs

(LIPIS and SSIM present the average evaluation values of same 100 SR results under different number of DRBs. Training time indicates different training time costs under the same training dataset.)

Additionally, in order to evaluate the effectiveness of the proposed components in TESRGAN, the comparison experiment shown in Fig. 12 was conducted by gradually adding components to the baseline SRGAN network. Each column depicts a module usage scheme, while the various reconstructions depict visual comparisons of various components. As can be seen, dense connection of residual blocks modifies the baseline's blurring situation. The CA module enhances the reconstruction results by focusing on the finer details. The edge extractor divides the LR image into high- and low-frequency images and feeds them into the GN, significantly improving the edge clarity and texture of the SR results. Due to the input of high-frequency information that divided by edge operator images, GN can devote more attention to reconstructing texture information.

We investigate the effect of various edge detection operators on SR and finally abandon the complex edge detection algorithm in favor of a simple convolution operator to align the model with the end-to-end ideology. The comparison experiments that employ first order convolution operator Sobel and second order convolution operator Laplace represented the reconstruction results produced by various edge operators do not differ significantly, and the Sobel operator is more robust to noise in comparison, which is why it is used in our framework.

Dense Connection	×	✓	✓	✓
Channel Attention	×	×	✓	✓
Edge Extractor	×	×	×	✓

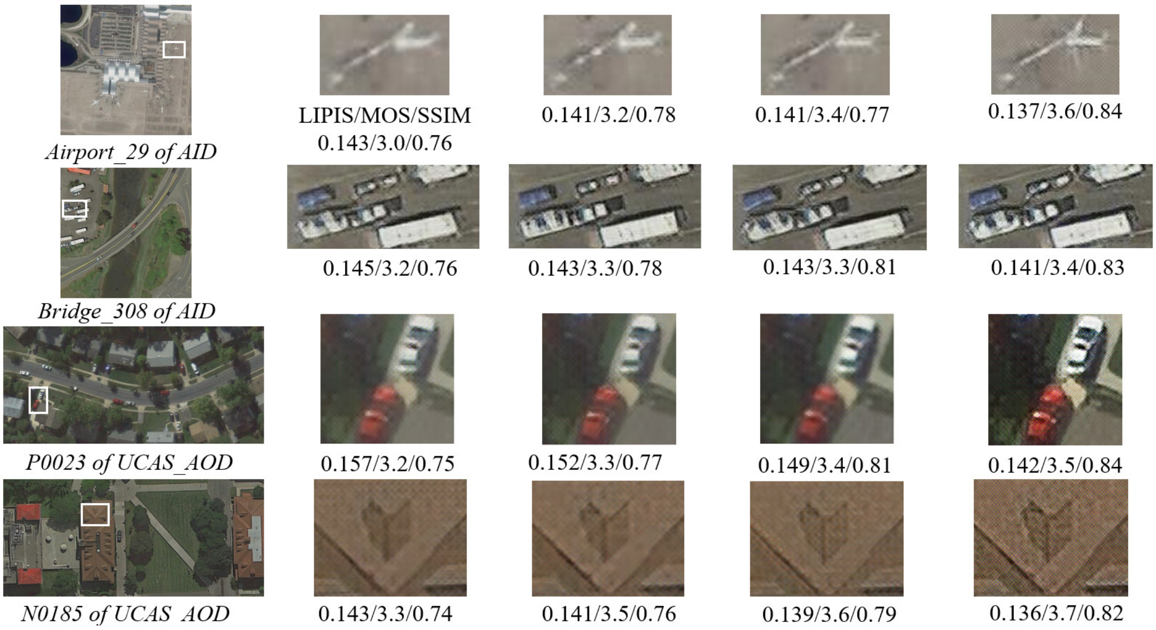


Fig. 12. Comparisons for presenting the effectiveness of each component in TESRGAN

5 Conclusion and Discussion

This paper proposes a new texture-enhanced generative adversarial network (TESRGAN) to super-resolution the LR remote sensing data. Specially, the proposed framework focuses on solving the problem that the SR results of previous algorithms are smooth and lack of details. A new DRB module with CA mechanism and dense connection was introduced to optimize the use of layer features and channel information. The improve structure based on the GAN baseline make full use and fusion of the information in different positions and channels, and avoid the gradient vanishing in training as much as possible. On the other hand, we introduce an edge extractor to divide the LR images that input directly to the network in previous methods into low and high frequency information, which can significantly improve the details description ability of SR results.

We compared the proposed TFSRGAN's SR results to those of other state-of-the-art CNN- and GAN-based SISR methods, including VDSR, SRGAN, and ESRGAN, in AID, UCAS_AOD and China Gaofen-1 datasets. The fourfold upscaling reconstruction comparison results indicated that our framework obtains a competitive global performance in terms of quantitative and qualitative evaluation. Regarding the metrics including LIPIS, SSIM and MOS, the SR results of the proposed method obtain the best performance on average. In the ground targets detection test, the SR results of proposed TFSRGAN is still superior than other methods, which verifies that TFSRGAN improves the post-processing effect of SR results.

Although the proposed method results are encouraging as a SR reconstruct model in remote sensing, the algorithm still has some disadvantages, including LR images sampling method, model training time cost control, etc. Our future work will be aimed to the following directions: 1) optimizing the LR data sampling method to make it more consistent with the natural degradation relationship between LR and HR remote sensing images. Meanwhile, considering the impact of noise. 2) extending the cost function to optimize the correctness and effectiveness of the error back propagation for model updating. 3) reducing the computational cost and time cost by designing new train strategies.

6 Acknowledgement

The authors are grateful for collaborative funding support from the Natural Science Foundation of Guizhou Province, China (No. QianKeHeZi [2021] 0445) and the Youth Program of the Education Foundation of Guizhou Province, China (No. QianJiaoHeKYZi [2020] 125).

References

[1] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sensing 12(10)(2020) 1662.

[2] T.-J. Feng, H.-R. Ma, X.-W. Cheng, Land-cover classification of high-resolution remote sensing image based on multi-classifier fusion and the improved Dempster–Shafer evidence theory, Journal of Applied Remote Sensing 15(1) (2021) 014506.

[3] H. Chen, W. Li, Z. Shi, Adversarial instance augmentation for building change detection in remote sensing images, IEEE Transactions on Geoscience and Remote Sensing 60(2021) 1-16.

[4] H.-S. Yue, J.-X. Cheng, Z. Liu, W. Chen, Remote-sensing image super-resolution using classifier-based generative adversarial networks, Journal of Applied Remote Sensing 14(4)(2020) 046514.

[5] X.-B. Feng, W.-X. Zhang, X.-Q. Su, Optical Remote Sensing Image Denoising and Super-Resolution Reconstructing Using Optimized Generative Network in Wavelet Transform Domain, Remote Sensing 13(9)(2020) 1858.

[6] Z. Pan, W. Ma, J. Guo, B. Lei, Super-resolution of single remote sensing image based on residual dense back projection networks, IEEE Transactions on Geoscience and Remote Sensing 57(10)(2019) 7918-7933.

[7] N. Zhang, Y. Wang, X. Zhang, D. Xu, X. Wang, An unsupervised remote sensing single-image super-resolution method based on generative adversarial network, IEEE Access 8(2020) 29027-29039.

[8] Y. Jo, S. Kim, Practical Single-Image Super-Resolution Using Look-Up Table, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[9] H. Chang, D.-Y. Yeung, Y. Xiong, Super-resolution Through Neighbor Embedding, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2004.

[10] F. Zhou, W. Yang, Q. Liao, Interpolation-based Image Super-resolution Using Multi-surface Fitting, IEEE Transactions on Image Processing 21(7)(2012) 3312-3318.

[11] C. Liu; D.-Q. Sun, On Bayesian Adaptive Video Super Resolution, IEEE Transactions on Pattern Analysis and Machine Intelligence 36(2)(2014) 346-360.

[12] V. Rahiman, S. George, Single Image Super Resolution Using Neighbor Embedding and Statistical Prediction Model, Computers and Electrical Engineering 62(2017) 281-292.

[13] C. Dong, C. Loy, K.-M. He, X. Tang, Image Super-resolution Using Deep Convolutional Network, IEEE Transactions on Pattern Analysis and Machine Intelligence 38(2)(2016) 295-307.

[14] C. Dong, C. Chen, K.-M. He, X. Tang, Learning a Deep Convolutional Network for Image Super-resolution, in: Proc. European Conference on Computer Vision (ECCV), 2014.

[15] W. Shi, J. Caballero, F. Huszár, J. Totz, A.P. Aitken, R. Bishop, D. Rueckert, Z. Wang, Real-time Single Image and Video Super-resolution Using an Efficient Sub-pixel Convolutional Neural Network, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[16] C. Dong, C. Loy, X. Tang, Accelerating the super-resolution convolutional neural network, in: Proc. European

Conference on Computer Vision (ECCV), 2016.

[17] K.-M. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proc. IEEE Conference on Com- puter Vision and Pattern Recognition (CVPR), 2016.

[18] J. Kim, J. Lee, K. Lee, Accurate Image Super-resolution Using Very Deep Convolutional Networks, in: Proc. IEEE Con ference on Computer Vision and Pattern Recognition (CVPR), 2016.

[19] B. Lim, S. Son, H. Kim, S. Nah, K. Lee, Enhanced Deep Residual Networks for Single Image Super-resolution, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[20] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, Y. Fu, Image Super-resolution Using Very Deep Residual Channel Attention Networks, in: Proc. European Conference on Computer Vision (ECCV), 2018.

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, W. Shi, Photo-realistic Single Image Super-Resolution Using a Generative Adversarial Network, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[22] X. Wan, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C.C. Loy, ESRGAN: Enhanced Super-resolution Generative Adversarial Networks, in: Proc. European Con ference on Computer Vision (ECCV), 2018.

[23] Y. Tian, R. Jia, S. Xu, R. Hua, M. Deng, Super-resolution Reconstruction of Remote Sensing Images Based on Convolutional Neural Network, Journal of Applied Remote Sensing 13(4)(2019) 046502.

[24] E. Maggiori, Y. Tarabalka, G. Charpiat, P. Alliez, Convolutional Neural Networks for Large-scale Remote-sensing Image Classification, IEEE Transactions on Geoscience and Remote Sensing 55(2)(2017) 645-657.

[25] P. Wang, H. Zhang, F. Zhou, Z. Jiang, Unsupervised remote sensing image super-resolution using cycle CNN, in: Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2019.

[26] Y. Xiong, S. Guo, J. Chen, X. Deng, L. Sun, X. Zheng, W. Xu, Improved SRGAN for remote sensing image super-resolution across locations and sensors, Remote Sensing 12(8)(2020) 1263.

[27] P. Wang, B. Bayram, E. Sertel, Super-resolution of Remotely Sensed Data Using Channel Attention Based Deep Learning Approach, International Journal of Remote Sensing 42(16)(2021) 6048-6065.

[28] S. Zhang, Q. Yuan, J. Li, J. Sun, X. Zhang, Scene-adaptive Remote Sensing Image Super-resolution Using a Multiscale Attention Network, IEEE Transactions on Geoscience and Remote Sensing 58(7)(2020) 4764-4779.

[29] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Nets, in: Proc. Advances in Neural Information Processing Systems (NIPS), 2014.

[30] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual Dense Network for Image Super-resolution, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[31] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation Networks, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[32] Y. Mei, Y. Fan, Y. Zhou, Image Super-Resolution with Non-Local Sparse Attention, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[33] Y. Blau, T. Michaeli, The Perception-distortion Tradeoff, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[34] R. Zhang, P. Isola, A. Efros, E. Shechtman, O. Wang, The Unreasonable Effectiveness of Deep Features as a Perceptual Metric, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.