DRGAN: A Detail Recovery-Based Model for Optical Remote Sensing Images Super-Resolution

Yongchao Song, Lijun Sun, Jiping Bi, Siwen Quan, and Xuan Wang, *Senior Member, IEEE*

***Abstract*— The need for high-resolution (HR) remote sensing images has grown significantly in recent years as a result of the rapid advancement of fine-sensing technologies. However, increasing sensor resolution usually requires a costly investment. To tackle this challenge, super-resolution (SR) methods for remote sensing images have emerged as a cost-effective alternative to enhance the quality and usability of existing low-resolution (LR) images. Although many current methods have achieved some reconstruction results, they often suffer from problems such as transition smoothing and artifacts. To solve these problems, we propose an SR reconstruction model for detail recovery based on generative adversarial networks (GANs), referred to as DRGAN. Specifically, unlike the traditional residual-in-residual dense block network (RRDBNet), we propose a novel dense residual network (OSRRDBNet). It uses dynamic convolution and self-attention mechanisms to recover the rich detailed information in the image more effectively. In addition, we employ an average pooling layer to enhance the ability to capture HR image features. By conducting experiments on three different remote sensing datasets, DRGAN shows remarkable reconstruction results and successfully recovers the rich detail information in the images.**

***Index Terms*— Detail recovery, dynamic convolution, generative adversarial network (GAN), optical remote sensing, super-resolution (SR).**

I. Introduction

With the increasing demand for fine remote sensing in various industries, the application scenarios of high-resolution (HR) remote sensing images are becoming increasingly extensive. Super-resolution (SR) image algorithmically enhances the quality and detail of LR image [1], [2], [3]. Currently, SR is receiving more and more attention and exploration in the field of remote sensing, including HR map generation [4], multimodal data fusion [5], and target detection [6].

The process of SR image is quite challenging. This is because the degradation of an image is not predictable, making it difficult to find an exact solution. Given an LR image, there may be countless HR versions, and the complexity makes image SR a daunting task. Therefore, it is crucial to research and develop an efficient method to improve the quality of remote sensing images, as it has significant practical applications.

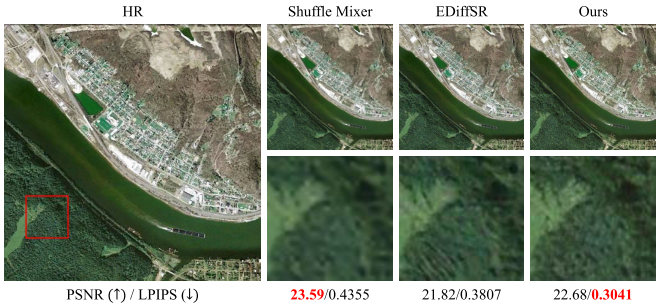

Fig. 1. Visual comparison of our DRGAN with the shuffle mixer and EDiffSR. DRGAN generates richer texture details that are more consistent with human eye perception.

Recent advancements in SR have been driven by deep learning-based methods [7], [8], [9], [10], [11]. While many of these methods aim for a high peak signal-to-noise ratio (PSNR), they often result in fuzzy, noisy outputs. For example, the shuffle mixer utilizes a large kernel convolution in a convolutional neural network (CNN), leading to the highest PSNR but a lack of texture detail, as illustrated in Fig. 1.

Many attempts have been made to enhance the perceived quality of images. For example, the CNN [12] and Transformer architectures [13], [14], as well as diffusion models [15], [16], have been employed. However, current methods often use local pixel smoothing techniques, like the mean square error loss function. These methods struggle to efficiently capture the complex structure and nonlocal pixel relationships in an image. While they may achieve a high PSNR, they often lead to overly smooth generated images that lack realism and detail.

Generative adversarial network (GAN) is a popular area of research in SR [17] and contains a generator and a discriminator. The generator is responsible for producing images with rich details. The discriminator evaluates the authenticity of the input image and thus helps the generator to improve the quality of the generated image. The GAN has been widely used to explore perception-oriented SR [18], [19], [20]. However, while these methods recover a certain level of image detail, they often introduce distortions and artifacts in the SR image. This means that the details produced by the generator are different from the real details.

Received 6 September 2024; revised 4 November 2024; accepted 2 December 2024. Date of publication 9 December 2024; date of current version 30 December 2024. This work was supported by the Natural Science Foundation of Shandong Province under Grant ZR2022QF037. *(Corresponding authors: Lijun Sun; Xuan Wang.)*

Yongchao Song, Lijun Sun, Jiping Bi, and Xuan Wang are with the School of Computer and Control Engineering, Yantai University, Yantai 264005, China (e-mail: ycsong@ytu.edu.cn; sunlijun@s.ytu.edu.cn; bijiping@s.ytu.edu.cn; xuanwang91@ytu.edu.cn).

Siwen Quan is with the School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China (e-mail: siwenquan@chd.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3512528

To address these issues, we propose a GAN-based detail recovery (DRGAN) SR model to enhance remote sensing images. This network can produce a pleasing effect that matches the perception of the human eye while maintaining authenticity and naturalness. Specifically, in DRGAN, the generator dynamic convolution and self-attention residual-in-residual dense block network (OSRRDBNet) has three important structures: dynamic convolution, self-attention mechanism, and dynamic convolution and self-attention residual dense block (OSRRDB). Dynamic convolution can dynamically adjust the shape and size of the convolution kernel based on the input data. This flexibility allows the model to accommodate features of different scales and structures better, especially when dealing with remote sensing images. The self-attention mechanism enhances the model's ability to perceive the global picture. The OSRRDB structure is also one of the key factors in its success. OSRRDB effectively enhances the model's ability to learn fine-grained features by stacking multiple residual connections and dense blocks. The introduction of this structure allows DRGAN to achieve higher image reconstruction quality and detail recovery while maintaining model compactness. The discriminator is inspired by the architecture of VGG19. An average pooling layer is utilized to enhance the ability to capture HR image features, encouraging the generator to generate images that are more similar to HR.

In summary, the main contributions of this article are at least threefold.

1) We propose a new SR model for remote sensing images based on detail recovery. By designing OSRRDBNet as the generator and adjusting the structure of the discriminator, the generation quality and evaluation of remote sensing images are significantly improved.
2) We designed the generator OSRRDBNet. OSRRDBNet consists of a self-attention mechanism, dynamic convolution, and OSRRDB, which significantly improves the accuracy of SR reconstruction of remote sensing images.
3) The proposed DRGAN outperforms previous SOTA methods in recovering different degrees of degradation, demonstrating robust reconstruction performance.

The rest of the article is organized as follows. Section II reviews the related work on single-image SR reconstruction. Section III presents detailed information references about the proposed DRGAN. Section IV provides extensive experimental results to validate the performance of DRGAN. Section V summarizes the full text.

II. Related Work

In this section, we provide a brief review of deep learning-based SR reconstruction methods, highlighting some reconstruction techniques and research advances.

A. CNN-Based Models

With the advancements in CNNs, there has been significant progress in CNN-based reconstruction methods. Inspired by the SRCNN [21], some researchers have proposed deeper and wider networks [22], [23]. Zhang et al. [24] used residual networks to address the gradient vanishing problem. While Liang et al. [25] focused on real-world image reconstruction using CNN-based multiexpert SR networks. Zhang et al. [26] introduced a closed-loop network framework for single-frame SR of infrared remote sensing images in real environments. Meanwhile, two generative CNNs were also used for downsampling and SR processing [27]. Additionally, Zhou et al. [28] developed a method for learning correction filters through degenerate adaptive regression modules in an unsupervised manner. However, it is worth noting that most CNN-based methods rely on pixel smoothing strategies, which, despite achieving high PSNR, often result in overly smooth outputs.

B. Transformer-Based Models

Due to its powerful self-attention mechanism, Transformers exhibit superior reconstruction capabilities when compared to CNNs. Lei et al. [29] introduced a multilevel augmented Transformer that uses self-attention to explore features at different scales, but it does not perform a global search at each stage. Liang et al. [30] achieved information sharing between neighboring pixels in single image SR (SISR) by introducing SwinIR with a shift window mechanism. He et al. [15] developed a dense spectral Transformer, incorporating ResNet, for multispectral remote sensing images to learn long-range relationships within the data. In addition, Xiao et al. [8] proposed an adaptive method to effectively eliminate the interference of irrelevant markers, thus making the self-attentive computation more efficient and compact.

C. Diffusion-Based Models

The recent acceleration of diffusion models has garnered significant attention. SR techniques based on diffusion modeling primarily generate HR images through a step-by-step denoising process. Luo et al. [31] proposed averaging equations to model image degradation while achieving a faster diffusion process. Xiao et al. [10] introduced an efficient activation network (EANet) along with a conditional prior augmentation module to simplify noise prediction and enhance computational efficiency. Similarly, Yue et al. [32] developed a Markov chain to facilitate the transfer between HR and low-resolution (LR) images, improving transfer efficiency. Although these acceleration methods for diffusion modeling have made progress in enhancing inference speed and computational efficiency, they still face limitations in adequately capturing complex image details and textures.

D. GAN-Based Models

GAN is one of the most promising methods for distributing unsupervised learning. The GAN model optimizes the performance of the generator and discriminator by adversarial training between these two components. Since its proposal, the GAN model has garnered extensive attention and research. Ledig et al. [18] utilize GANs to achieve photo-level SR of photo-realistic sensory single images. Li et al. [33] utilized a region-aware adversarial learning strategy to instruct the

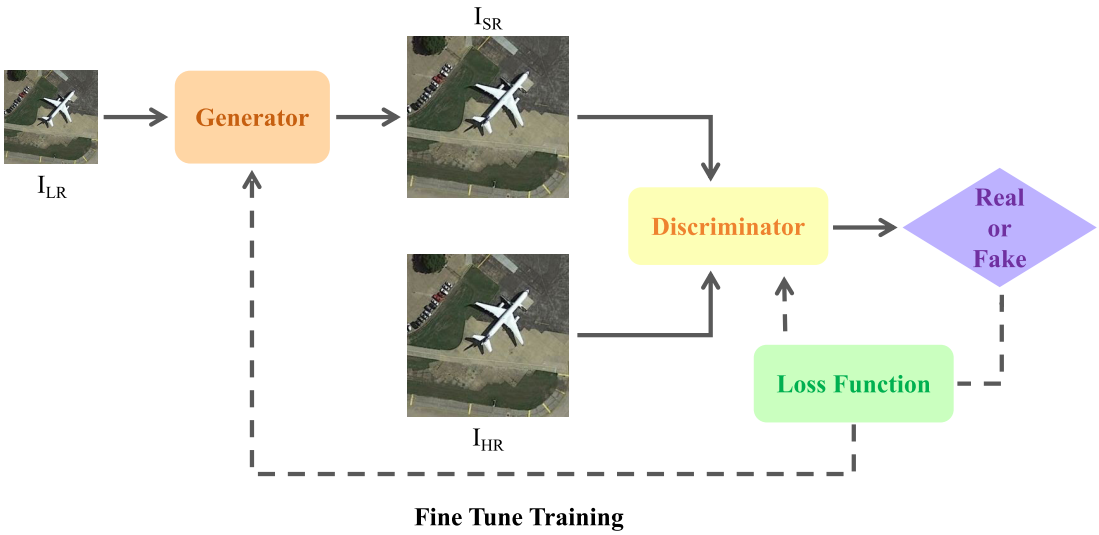

Fig. 2. Overall framework of the proposed DRGAN.

model to focus on the details of adaptively generated texture regions. However, the robustness of the model is not satisfactory and the color fidelity of the generated images is low. Dong et al. [34] explored the potential of a reference-based SR (RefSR) approach in remote-aware images to reconstruct the details of LR images using rich texture information from HR reference images. Furthermore, Li et al. [35] proposed a semantic-aware discriminator to improve the GAN by introducing semantic information to produce closer-to-real image results. In contrast, DRGAN not only emphasizes pixel-level details but also effectively captures complex textures and structural information in images by incorporating a self-attention mechanism and dynamic convolution. As a result, there is a significant improvement in color accuracy and detail reproduction in the generated images. Additionally, the design of DRGAN prioritizes the model's robustness and minimizes artifacts in the generated images, thereby enhancing their realism.

III. Methodology

A. Overview

In this article, we aim to design a model suitable for SR reconstruction of optical remote sensing images. Our proposed DRGAN framework, as depicted in Fig. 2, consists of three main components: an LR image input, a generator (G), and a discriminator (D). The generator can be roughly divided into three substructures: the OSRRDB, the convolutional layer, and the upsampling. First, the input image is subjected to feature extraction. Subsequently, the input features are extracted and reconstructed at a deep level using OSRRDB. Finally, the reconstructed image is generated by further processing and magnification through convolutional and upsampling layers. The process can be expressed as follows:

$$Y = f_{3\times3}(f_{3\times3}(f_{\text{up}}(f_{3\times3}(f_{\text{os}}(f_{3\times3}(X)))))) \tag{1}$$

where Y is the reconstructed HR image and X is the input LR image. $f_{3\times3}$ denotes a 3×3 convolution and f_{up} operation for upsampling operation. f_{os} denotes processing by the OSRRDB module.

During training, the generator attempts to transform the input LR image into a space that closely resembles the real HR image, thereby creating a reconstructed image. For the discriminator, we designed the precision extractor VGG19 (PEVGG19) inspired by the architecture of VGG19. Instead of using the maximum pooling layer, an average pooling layer is used, which motivates the generator to produce images that are more similar to the HR. Two metrics, PSNR and LPIPS, are oriented, and G and D are updated by iterative training and updating until convergence to SR is similar to HR.

B. Generator Network of DRGAN

As shown in Fig. 3, the main component of our generator network, OSRRDBNet, is the dynamic dense residual block OSRRDB. With residual connections, the model efficiently learns and reconstructs the input image details and features.

Fig. 3 illustrates the architecture of OSRRDB. Each OSRRDB consists of an OSRDB connected with a self-attention mechanism. The working process of OSRRDB can be summarized as follows:

$$Y_n = \text{OSRRDB}_n(X_{n-1}) + X_{n-1} \tag{2}$$

where Y_n is the output of the nth OSRRDB, X_{n-1} is the input of the previous layer (or the initial input (X)), and OSRRDB_n is the operation of the nth OSRRDB.

The OSRDB is primarily made up of dynamic convolution and LeakyReLU. Dynamic convolution and activation units are linked through residual connections, enabling previous layer inputs and outputs to be fully utilized within each dynamic dense residual block. The OSRDB design enables feature reuse and information transfer, enhancing feature extraction. In contrast to RRDB architectures containing 3×3 convolution and activation layers, this design can efficiently capture long-term dependencies and important features in images.

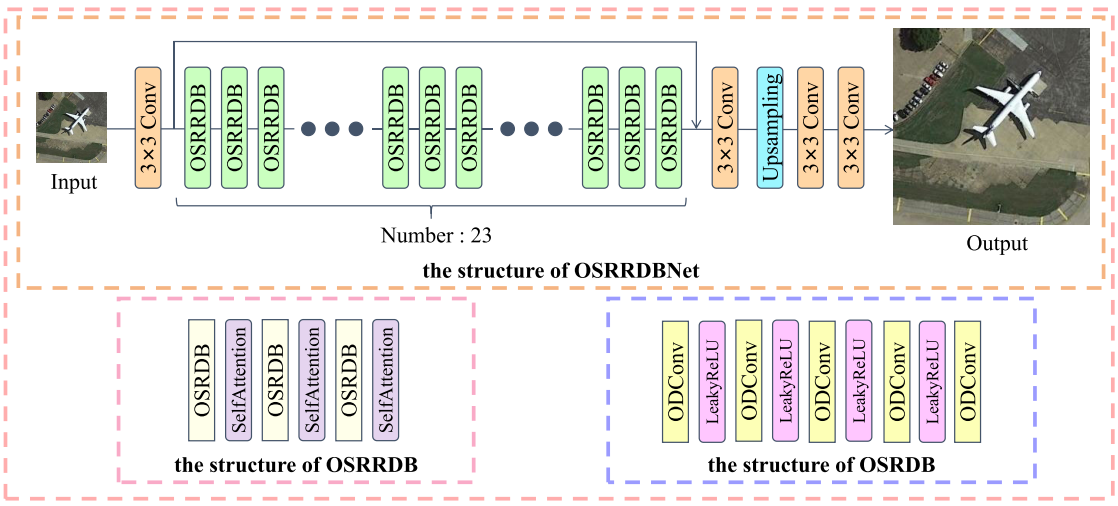

Fig. 3. Framework of the generator (OSRRDBNet) for DRGAN. Among them, OSRRDB consists of an OSRDB and a self-attention mechanism. OSRDB consists of dynamic convolution and LeakyReLU.

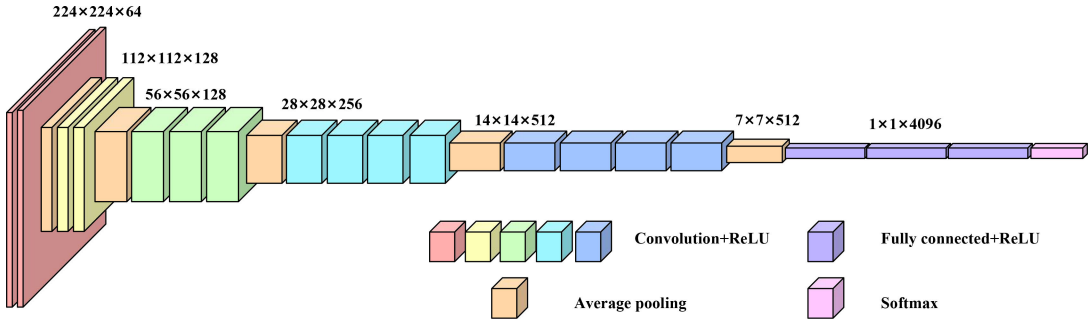

Fig. 4. Framework of the discriminator (PEVGG19) for DRGAN.

Dynamic convolution is a key feature of OSRRDB. Unlike traditional 3 × 3 convolution, dynamic convolution can adjust the shape and parameters of the convolution kernel based on the input data characteristics. This enables the network to adapt better to different types and scales of features, leading to more efficient capture of details and structures in the image.

The self-attention mechanism used in OSRRDB is a lightweight attention mechanism for enhancing the model's ability to perceive the global picture. By introducing self-attention in each OSRRDB, the network can better understand the dependencies between the parts in the image, which helps to improve the accuracy and quality of image reconstruction.

Compared to traditional ReLU, LeakyReLU introduces a small negative slope when dealing with the activation function. This makes the network more stable during training and able to handle a wider range of input distributions.

Our OSRRDB allows the model to perform exceptionally well in enhancing the resolution of optical remote sensing images. This is achieved through its distinctive structural design, which incorporates residual connectivity, dense blocks, and self-attention mechanisms. It not only accurately captures long-term dependencies but also extracts and retains crucial features in the image. As a result, this significantly enhances the quality and visual appeal of the final reconstructed image.

C. Discriminator Network of DRGAN

As shown in Fig. 4, our discriminator employs an average pooling layer to enhance its ability to capture HR image features, prompting the generator to produce images that are more similar to the HR.

Three main advantages of using an average pooling layer over a maximum pooling layer in the discriminator are as follows.

1) *Smooth Feature Extraction:* The average pooling layer calculates the average of all the pixels in a specific area, creating a more consistent feature representation. This helps to smooth out noise and variations in the feature map, which is crucial for capturing the overall structure and texture of an image. It also assists the discriminator in accurately distinguishing between an HR image and a generated image.
2) *Enhanced Feature Stability:* Unlike the maximum pooling layer, the average pooling layer does not lose any pixel information in the pooled region, but instead combines the average of all pixels in the region. It makes the discriminator more robust to changes in image features, especially when dealing with images with large variations and noise. By ensuring that the discriminator accurately evaluates the fidelity and quality of the

Fig. 5. Visual comparison with the SOTA SR model on the AID test set. The results show that our DRGAN significantly outperforms the comparison method in terms of high-frequency detail recovery.

TABLE I

QUANTITATIVE COMPARISON OF FID AND LPIPS WITH SOTA SR MODELS IN THE 30 SCENARIO CATEGORIES OF THE AID TEST SET. THE BEST VALUES IN EACH CATEGORY ARE HIGHLIGHTED IN RED, AND THE SECOND-BEST VALUES ARE HIGHLIGHTED IN BLUE

Categories	Metrics	Bicubic	SRGAN	ESRGAN	BSRGAN	Beby-GAN	DRSR	ShuffleMixer	EDiffSR	Ours
Airport	LPIPS	0.4796	0.3606	0.2387	0.3491	0.8879	0.3828	0.3353	0.2866	0.2245
	FID	76.02	86.31	46.54	115.26	346.56	79.87	72.20	69.78	46.62
BareLand	LPIPS	0.3704	0.3125	0.2919	0.3895	0.9008	0.3573	0.3555	0.3233	0.2671
	FID	67.69	88.42	70.12	142.15	346.24	103.73	75.64	71.49	64.73
BaseballField	LPIPS	0.3891	0.2947	0.2274	0.2908	0.8961	0.3093	0.3036	0.2382	0.2042
	FID	86.35	99.77	51.03	99.06	326.58	82.16	65.35	65.44	49.43
Beach	LPIPS	0.4151	0.3712	0.2868	0.4221	0.8026	0.3773	0.3740	0.3324	0.2613
	FID	61.97	85.82	57.05	134.42	316.18	79.33	66.11	65.72	46.86
Bridge	LPIPS	0.4306	0.3554	0.2453	0.3394	0.7985	0.3541	0.3412	0.2876	0.2367
	FID	78.43	86.01	39.94	96.83	336.47	73.26	57.40	59.41	38.90
Center	LPIPS	0.4893	0.3740	0.2367	0.3536	0.8920	0.3851	0.3210	0.2789	0.2252
	FID	80.69	85.07	38.53	92.27	307.97	72.10	58.24	62.97	37.32
Church	LPIPS	0.5568	0.4278	0.2707	0.3878	0.9126	0.4463	0.3738	0.3125	0.2605
	FID	84.87	100.61	47.32	112.28	367.51	86.33	76.53	71.78	45.30
Commercial	LPIPS	0.4684	0.3466	0.2411	0.3602	0.9627	0.3716	0.3278	0.2845	0.2279
	FID	64.56	97.98	37.58	139.38	448.71	93.39	62.47	69.30	39.55
DenseResidential	LPIPS	0.5821	0.4460	0.3053	0.3889	0.9928	0.4829	0.4231	0.3349	0.2962
	FID	77.21	92.44	36.78	94.99	402.69	94.53	69.76	62.57	33.04
Desert	LPIPS	0.3578	0.2958	0.2131	0.3146	0.8748	0.3206	0.3385	0.2556	0.196
	FID	69.63	164.04	56.82	152.62	354.12	82.19	67.63	67.02	57.37
Farmland	LPIPS	0.4245	0.3447	0.2392	0.3426	0.8351	0.3478	0.3290	0.2865	0.2233
	FID	78.26	86.23	42.29	99.54	295.49	76.99	61.49	61.36	44.58
Forest	LPIPS	0.5002	0.4438	0.3973	0.6233	0.8392	0.5167	0.4622	0.3953	0.306
	FID	64.38	82.33	79.72	164.56	300.48	88.74	74.49	96.94	53.73
Industrial	LPIPS	0.4766	0.3553	0.2291	0.3488	0.9442	0.3770	0.3182	0.2747	0.2139
	FID	62.89	79.13	32.02	107.78	368.34	71.88	57.95	65.19	31.31
Meadow	LPIPS	0.4867	0.4338	0.3710	0.4719	0.6499	0.4808	0.4832	0.3923	0.3085
	FID	83.96	94.99	65.84	111.74	278.07	98.91	78.91	80.59	53.94
MediumResidential	LPIPS	0.5056	0.3991	0.2866	0.3812	0.9082	0.4321	0.3867	0.3006	0.2596
	FID	86.10	94.25	44.33	91.82	402.55	86.82	73.15	66.80	39.48
Mountain	LPIPS	0.5210	0.4140	0.3649	0.3982	0.8636	0.4542	0.4449	0.3356	0.2985
	FID	67.89	86.64	61.39	140.42	331.61	86.89	61.51	55.51	58.34
Park	LPIPS	0.4884	0.3755	0.2943	0.3924	0.9286	0.4050	0.3782	0.3154	0.2547
	FID	70.11	88.43	52.45	139.03	402.41	81.64	65.77	76.34	48.80
Parking	LPIPS	0.3756	0.2615	0.1857	0.2844	0.9760	0.2846	0.2289	0.2126	0.1716
	FID	69.61	70.53	29.40	76.41	345.41	58.48	43.00	52.03	25.37
Playground	LPIPS	0.4007	0.3237	0.2130	0.3176	0.8655	0.3212	0.2931	0.2417	0.1981
	FID	72.68	76.93	34.64	84.69	304.75	66.06	51.09	52.77	33.68
Pond	LPIPS	0.4581	0.3881	0.2871	0.3609	0.8052	0.4046	0.3944	0.3359	0.2624
	FID	79.87	92.30	51.26	108.91	307.41	90.98	66.23	68.69	46.24
Port	LPIPS	0.4343	0.3475	0.2242	0.3169	0.8578	0.3547	0.3167	0.2809	0.2133
	FID	76.98	92.14	38.27	95.74	362.89	72.27	53.99	63.68	38.86
RailwayStation	LPIPS	0.484	0.3653	0.2397	0.3645	0.9270	0.3761	0.3187	0.2992	0.2181
	FID	66.35	84.33	39.86	149.36	399.16	82.98	67.11	89.48	39.60
Resort	LPIPS	0.488	0.3758	0.2805	0.3777	0.9253	0.4047	0.3695	0.3034	0.256
	FID	78.85	104.46	51.35	120.19	390.78	83.47	67.29	72.17	47.65
River	LPIPS	0.4816	0.3978	0.3152	0.3886	0.8262	0.4321	0.4074	0.3312	0.2562
	FID	78.92	88.99	51.96	105.14	295.27	85.67	62.37	63.68	49.39
School	LPIPS	0.5065	0.3829	0.2620	0.3624	0.9532	0.4088	0.3635	0.2910	0.2468
	FID	60.71	77.23	37.40	105.77	409.45	75.58	58.85	60.11	36.46
SparseResidential	LPIPS	0.5699	0.4924	0.3887	0.4683	0.8299	0.5468	0.5184	0.3943	0.3239
	FID	90.10	110.76	55.98	89.57	361.62	89.43	70.26	75.33	49.60
Square	LPIPS	0.4414	0.3330	0.2412	0.3470	0.9167	0.3602	0.3151	0.2613	0.2207
	FID	61.39	69.72	35.96	90.26	286.67	64.25	52.31	49.05	32.86
Stadium	LPIPS	0.4648	0.3523	0.2321	0.3350	0.8848	0.3679	0.3122	0.2727	0.2214
	FID	70.99	77.27	30.72	79.35	299.12	59.54	46.22	54.09	30.18
StorageTanks	LPIPS	0.5355	0.4146	0.2708	0.3669	0.8920	0.4322	0.3816	0.3109	0.2633
	FID	99.13	97.16	41.91	100.07	330.49	82.70	64.43	64.51	41.16
Viaduct	LPIPS	0.4962	0.3857	0.2517	0.3479	0.8916	0.3889	0.3393	0.2813	0.2271
	FID	57.16	60.04	26.39	87.74	304.77	56.67	47.76	42.66	25.27
Average	LPIPS	0.4693	0.3724	0.2710	0.3731	0.8814	0.3961	0.3618	0.3017	0.2448
	FID	74.13	90.01	46.16	110.91	344.33	80.23	63.18	65.88	42.85

TABLE II
AVERAGE PSNR RESULTS FOR DIFFERENT MODELS RECONSTRUCTED ON THE WHU-RS19 DATASET. THE BEST VALUES IN EACH CATEGORY ARE HIGHLIGHTED IN RED, AND THE SECOND-BEST VALUES ARE HIGHLIGHTED IN BLUE

Methods \ kernel size	0.6	1.2	1.8
Bicubic	19.23	19.39	19.43
SRGAN	24.64	25.16	24.74
ESRGAN	24.95	25.96	25.79
BSRGAN	23.77	23.96	24.01
ShuffleMixer	28.03	28.17	27.16
EDiffSR	25.16	26.14	26.01
Ours	25.91	27.02	26.55

generated images, the average pooling layer guides the generator to produce more realistic and detailed SR images.

3) *Low Deviation:* The average pooling layer alleviates the over-reliance on individual pixel values in the feature map, which reduces the risk of biasing performance due to the use of max pooling. It helps to retain the global information present in the feature map, offering the discriminator the ability to more precisely assess the quality of the generated image for similarity to the HR image.

D. Training Optimization Strategies

During the training process, iterative optimization is used to update the generator and discriminator alternately. The discriminator gives feedback to the generator about image quality by comparing the generated SR image with the real HR image, which helps the generator continuously optimize the generation process. However, conventional evaluation metrics such as the PSNR and structural similarity index (SSIM) used in previous studies may struggle to accurately assess the perceived image quality in generated images.

To better evaluate the visual quality of the generated images, we introduce the learned perceptual image patch similarity (LPIPS) metric, an efficient model that simulates human perception. The similarity between SR and HR images generated is regularly evaluated by monitoring changes in the model's PSNR and LPIPS metrics on the validation set. This ensures that the generator produces high-quality images that are closer to the real images.

With this training optimization strategy, not only is the sensitivity of traditional metrics to numerical errors taken into account, but also a deeper understanding of the perceived quality of the incorporated image. It allows us to evaluate the generator's performance more comprehensively and further optimize the training process to obtain more realistic SR images.

IV. EXPERIMENTS AND DISCUSSIONS

In this section, we conduct extensive experiments on three datasets to evaluate the performance of DRGAN in different degradation scenarios.

A. Dataset

The used datasets for training are the ITCVD dataset [36] and the DLR Munich vehicle dataset [37]. They contain a total of 135 images. We use three publicly available remote sensing datasets, including AID, WHU-RS19 [38], and NWPU-RESISC 45 [39], to evaluate our approach. The AID dataset is a widely used benchmark for testing and assessing the performance of various computer vision algorithms in processing HR remote sensing images. Specifically, it contains different categories of remote sensing images, such as buildings, farmland, roads, and so on. The image sources in AID are varied and come from various sensors. It contains a variety of remote sensing images, including buildings, farmland, roads, and more, sourced from a range of sensors. The AID dataset contains 10 000 images, each with a resolution of 600 × 600 pixels. We use it to evaluate our approach in eight different ways.

Furthermore, WHU-RS19 includes soaring-resolution remote sensing imagery from assorted geographic environments, encompassing 19 distinct feature categories, such as buildings, forests, and water bodies. This dataset encapsulates HR, multifaceted feature classes, and authentic geography, rendering it a treasure trove of experimental value and potential applications. In our study, each of these 19 categories was degraded one at a time to three degenerate nuclei of varying sizes to obtain LR. Subsequently, the LR in this article model was compared with seven different reconstruction models.

The NWPU-RESISC 45 dataset was exclusively used for real-world analysis, without any modeling deterioration. To conserve inference cost, we randomly selected 135 images from the dataset, encompassing all 45 categories.

B. Evaluation Metrics

In this article, we mainly use six evaluation metrics to comprehensively evaluate the performance of the proposed method kernel and other comparative methods. The PSNR and SSIM [40] are the most frequently used parameters for image SR. The PSNR evaluates image quality primarily based on mean square error between image pixels but is not sensitive to human eye perception. SSIM evaluates image quality by comparing the similarity of structural information, including brightness, contrast, and structure. However, SSIM's assessment of an image is based on a localized region, not a global one. The third metric is LPIPS [41], which is a perceptual image similarity metric based on deep learning. Unlike the traditional PSNR and SSIM, LPIPS can more accurately capture differences in image quality as perceived by the human eye.

Frechet inception distance (FID) [42] is a metric used to evaluate the quality of generative models, especially widely used in the GAN. It measures the quality of the generated image by comparing the distance between the feature distribution of the generated image and the feature distribution of the real image. Additionally, we utilized natural image quality evaluator (NIQE) [43] and average gradient (AG), two reference-free image quality assessment methods, to evaluate

TABLE III

QUANTITATIVE COMPARISON WITH THE SOTA SR MODELS IN TERMS OF LPIPS ON THE WHU-RS19 TEST SET. THE BEST VALUES IN EACH CATEGORY ARE HIGHLIGHTED IN RED, AND THE SECOND-BEST VALUES ARE HIGHLIGHTED IN BLUE

	Bicubic	SRGAN	ESRGAN	BSRGAN	ShuffleMixer	EDiffSR	Ours
kernel size	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Airport	0.6308	0.3129	0.2091	0.3085	0.2774	0.2595	0.1871
Beach	0.5917	0.1835	0.2877	0.3965	0.1746	0.2832	0.1492
Bridge	0.6392	0.2520	0.2010	0.3095	0.2545	0.2611	0.1772
Commercial	0.5125	0.3776	0.2511	0.3518	0.3469	0.2947	0.2360
Desert	0.8825	0.3110	0.2176	0.3172	0.3535	0.2799	0.2161
Farmland	0.7043	0.2874	0.2756	0.3925	0.3005	0.2764	0.2201
footballField	0.5251	0.2911	0.1986	0.3019	0.2552	0.2337	0.1813
Forest	0.4831	0.4293	0.3533	0.5592	0.4326	0.3624	0.2937
Industrial	0.5202	0.3096	0.2103	0.3184	0.2785	0.2609	0.1873
Meadow	0.7018	0.3185	0.2749	0.3517	0.3562	0.2557	0.2116
Mountain	0.6450	0.5314	0.4283	0.4932	0.5396	0.4168	0.3898
Park	0.5870	0.3950	0.3000	0.4218	0.3895	0.3438	0.2654
Parking	0.4993	0.2250	0.1860	0.2611	0.1901	0.2130	0.1681
Pond	0.6131	0.2853	0.2320	0.2853	0.3039	0.2584	0.1900
Port	0.5159	0.2931	0.1968	0.2869	0.2636	0.2430	0.1790
RailwayStation	0.6315	0.3771	0.2306	0.3722	0.3107	0.3392	0.2146
Residential	0.4836	0.3278	0.2244	0.3091	0.3003	0.2577	0.2090
River	0.6051	0.3900	0.3141	0.3838	0.4113	0.3263	0.2530
Viaduct	0.5451	0.3522	0.2039	0.3079	0.2790	0.2546	0.1890
Average	0.5956	0.3289	0.2524	0.3541	0.3167	0.2853	0.2167
kernel size	1.2	1.2	1.2	1.2	1.2	1.2	1.2
Airport	0.6880	0.3031	0.2257	0.3070	0.3101	0.2523	0.2026
Beach	0.5991	0.1978	0.2745	0.3968	0.1840	0.2726	0.1390
Bridge	0.6718	0.2613	0.2083	0.3102	0.2871	0.2646	0.1902
Commercial	0.5664	0.3620	0.2529	0.3476	0.3709	0.2886	0.2384
Desert	0.8870	0.3428	0.2173	0.3162	0.3982	0.2987	0.2580
Farmland	0.7495	0.2894	0.2766	0.3900	0.3510	0.2938	0.2368
footballField	0.5721	0.2905	0.2046	0.2991	0.2854	0.2395	0.1916
Forest	0.5943	0.4252	0.3842	0.5553	0.4615	0.3731	0.2954
Industrial	0.5872	0.3020	0.2155	0.3156	0.3059	0.2531	0.1972
Meadow	0.7518	0.3108	0.2683	0.3501	0.4146	0.2801	0.2216
Mountain	0.7055	0.5279	0.4403	0.4907	0.5658	0.4223	0.3839
Park	0.6564	0.3860	0.3082	0.4196	0.4204	0.3407	0.2715
Parking	0.5590	0.2288	0.1865	0.2611	0.2147	0.2146	0.1645
Pond	0.6576	0.2841	0.2356	0.2851	0.3319	0.2688	0.2025
Port	0.5630	0.2986	0.2017	0.2856	0.2886	0.2494	0.1866
RailwayStation	0.6950	0.3723	0.2606	0.3727	0.3498	0.3118	0.2351
Residential	0.5437	0.3168	0.2250	0.3039	0.3214	0.2533	0.2061
River	0.6790	0.3774	0.3184	0.3813	0.4424	0.3277	0.2612
Viaduct	0.5994	0.3254	0.2207	0.3061	0.3116	0.2480	0.2055
Average	0.6487	0.3264	0.2592	0.3523	0.3482	0.2870	0.2257
kernel size	1.8	1.8	1.8	1.8	1.8	1.8	1.8
Airport	0.7619	0.3574	0.3248	0.3078	0.3947	0.3247	0.2988
Beach	0.6066	0.2234	0.2666	0.3974	0.2006	0.2728	0.1453
Bridge	0.7094	0.3076	0.2608	0.3113	0.3416	0.3070	0.2498
Commercial	0.6584	0.4042	0.3559	0.3461	0.4527	0.3653	0.3384
Desert	0.8893	0.4014	0.2497	0.3181	0.4441	0.3301	0.3168
Farmland	0.7932	0.3521	0.3397	0.3889	0.4388	0.3531	0.3083
footballField	0.6338	0.3431	0.2770	0.2992	0.3628	0.2980	0.2680
Forest	0.7312	0.4807	0.4226	0.5549	0.5525	0.4500	0.3724
Industrial	0.6791	0.3538	0.2956	0.3136	0.3874	0.3138	0.2866
Meadow	0.7887	0.3691	0.3214	0.3483	0.4903	0.3395	0.2914
Mountain	0.7936	0.5531	0.4984	0.4932	0.6332	0.4881	0.4594
Park	0.7471	0.4333	0.3882	0.4198	0.5039	0.4198	0.3735
Parking	0.6495	0.2899	0.2488	0.2669	0.2953	0.2675	0.2287
Pond	0.7123	0.3208	0.2769	0.2857	0.3871	0.3145	0.2591
Port	0.6315	0.3490	0.2665	0.2878	0.3574	0.3016	0.2580
RailwayStation	0.7791	0.4432	0.3774	0.3771	0.4478	0.3725	0.3585
Residential	0.6441	0.3602	0.3106	0.3025	0.4016	0.3189	0.2893
River	0.7710	0.4214	0.3730	0.3818	0.5208	0.3932	0.3490
Viaduct	0.6883	0.3663	0.3327	0.3085	0.4079	0.3262	0.3164
Average	0.7194	0.3753	0.3256	0.3531	0.4221	0.3451	0.3036

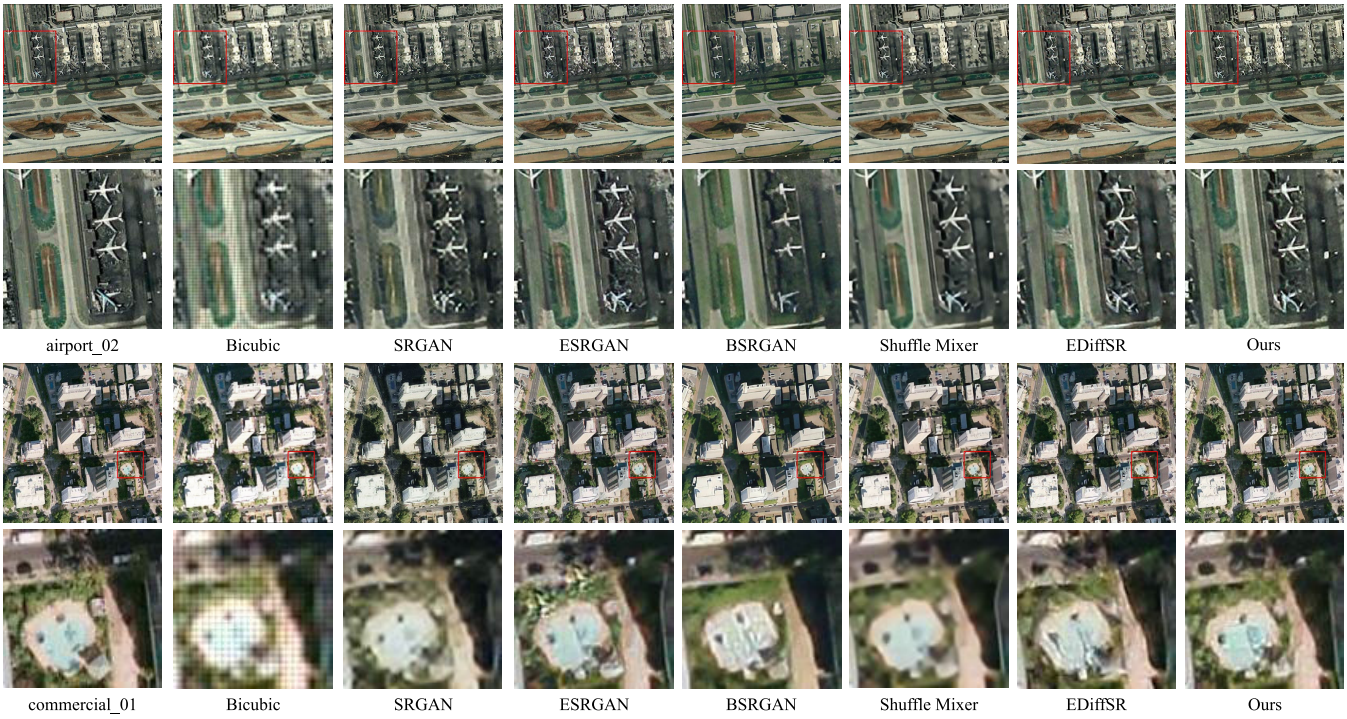

Fig. 6. Visual comparison with the SOTA SR model on the WHU-RS19 test set with a degeneracy kernel of 0.6. Local zoom better demonstrates image details.

real-world remote sensing images without the need for HR images.

C. Implementation Details

This study focuses on the ×4 SR. The number of OSRRDBs in the model is 23. The model is optimized using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. To ensure stability, a two-stage training strategy is used. First, we train the PSNR-oriented generator with the learning rate set to 2×10^{-4}. Subsequently, the entire network is trained with 1×10^{-4} as the initial learning rate. A total of 400 000 iterations are performed, with the learning rate decaying to half of the previous value every 50 000 iterations. All SR methods covered in this article are retrained on the same training set. Our experiments are implemented on 2× NVIDIA RTX 3090 GPUs.

D. Comparison With SOTAs

We compare our DRGAN with state-of-the-art SR methods, including Bicubic, SRGAN [18], ESRGAN [19], BSRGAN [20], Beby-GAN [33], DRSR [1], ShuffleMixer [7], and EDiffSR [10]. The methods we have chosen for these comparisons are representative of the mainstream methods in the field. We retrained these comparison methods on the same training set according to the official experimental details.

1) Quantitative Comparison: Table I provides the results of the quantitative comparison of FID and LPIPS on the AID dataset. We compute metrics for different types of scenarios and give average results for each method on the dataset. The results show that our proposed DRGAN obtains the best scores in both metrics. It is worth noting that achieving perfect reconstruction results is still a challenging task due to the complexity and variability of scenes in remote sensing images. Specifically, DRGAN has an average advantage of 3.31 over the second-best method (ESRGAN) in terms of FID. In terms of LPIPS, DRGAN has a significant reduction compared to other methods. The results show that the DRGAN algorithm can provide stable detail retrieval under different remote sensing scenarios, demonstrating good generative performance.

We downsampled images from the WHU-RS19 dataset at three different sizes of Gaussian fuzzy kernels (0.6, 1.2, and 1.8) and then tested the reconstruction performance of each model. We assess the PSNR values for the WHU-RS19 dataset in more detail in Table II. The table shows that our DRGAN achieves the next best value of the PSNR. ShuffleMixer achieves a high PSNR but the visual perception of the image falls short of expectations by minimizing L1 and frequency losses.

Moreover, Table III displays the LPIPS averages for the WHU-RS19 dataset. We observe that our DRGAN achieves the best LPIPS performance with degenerate cores of different sizes. For example, compared with EDiffSR, DRGAN reduces 24.01%, 21.39%, and 12.02% for Gaussian kernels of 0.6, 1.2, and 1.8, respectively. This demonstrates the strong generalization ability of our method on different degenerate kernels.

2) Qualitative Comparison: To better visualize the comparative results of the different models, the reconstruction results of the various methods for the four scenarios are shown in Fig. 5. We process images from the AID dataset uniformly using bicubic degradation and then input them into individual models for reconstruction. From the figure, we can see that our

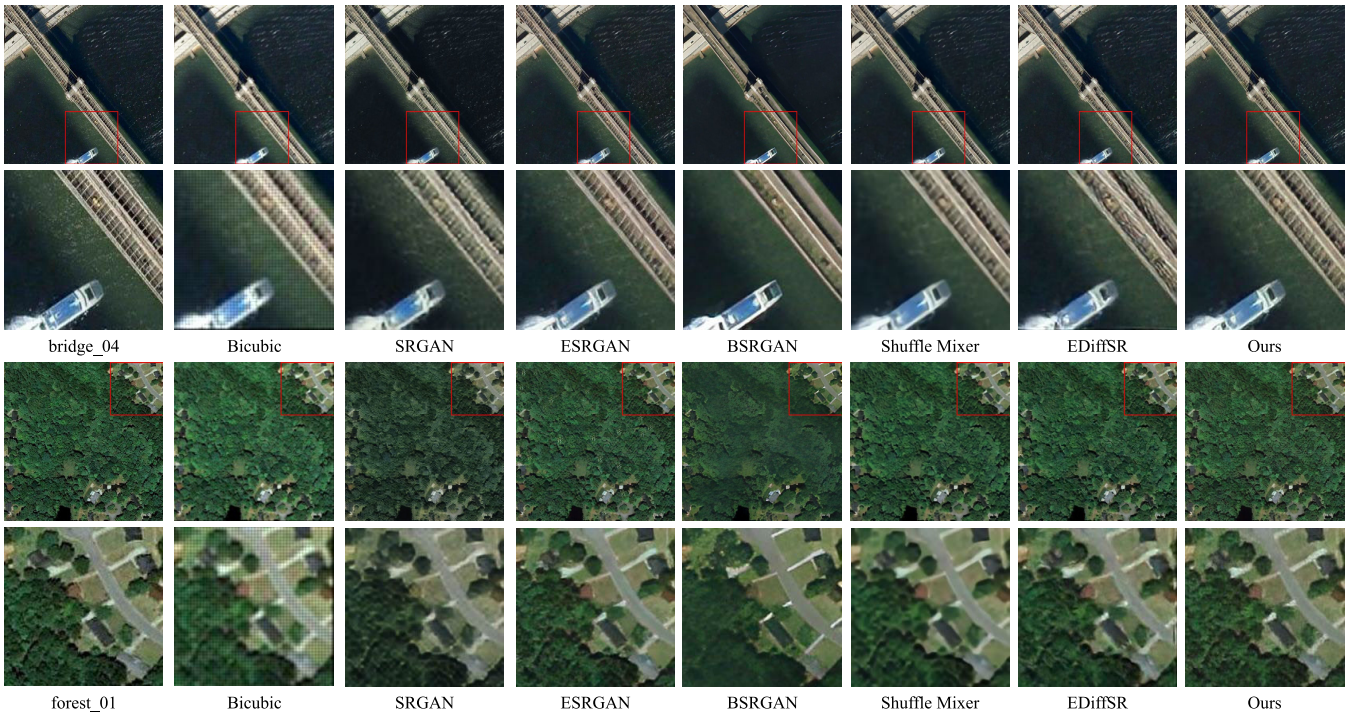

Fig. 7. Visual comparison with the SOTA SR model on the WHU-RS19 test set with a degeneracy kernel of 1.2. Local zoom better demonstrates image details.

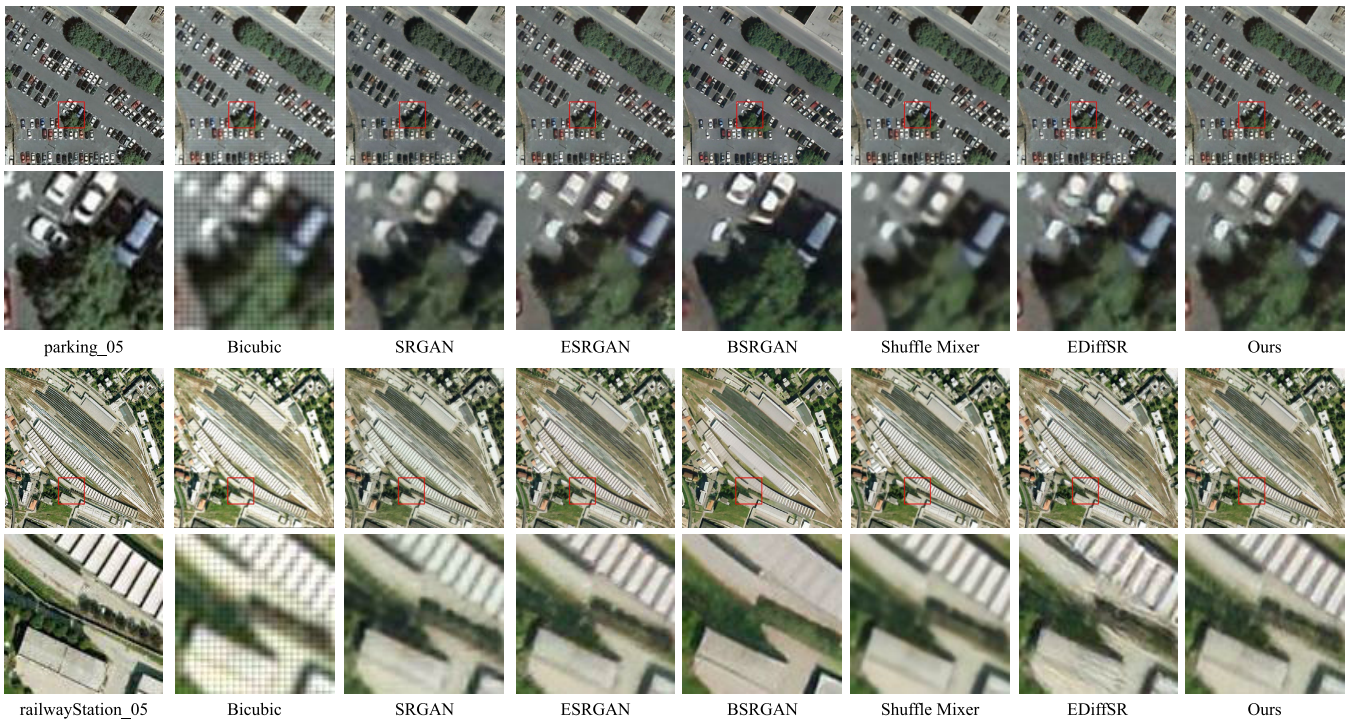

Fig. 8. Visual comparison with the SOTA SR model on the WHU-RS19 test set with a degeneracy kernel of 1.8. Local zoom better demonstrates image details.

DRGAN can consistently produce realistic results that outperform the SOTA methods. For the baseball_field151, the results produced by Beby-GAN are severely distorted, highlighting the limitations in preserving fine knots and clarity. In contrast, DRGAN provides a more natural effect. In playground_184, ESRGAN and ShuffleMixer can provide relatively realistic distributions. But there are still some transition smoothing issues that blur the details. In contrast, DRGAN can accurately generate details and present a more natural perception. These results highlight the generative power of OSRRDBNet,

TABLE IV

QUANTITATIVE COMPARISON WITH THE SOTA SR MODEL FOR NIQE AND AG. THE BEST VALUES IN EACH CATEGORY ARE HIGHLIGHTED IN RED, AND THE SECOND-BEST VALUES ARE HIGHLIGHTED IN BLUE

	Bicubic	SRGAN	ESRGAN	Beby-GAN	BSRGAN	TTST	EDiffSR	Ours
NIQE	20.806	15.098	16.887	23.336	14.497	14.590	17.932	4.9711
AG	6.924	2.770	3.594	6.219	3.229	3.100	3.323	3.3594

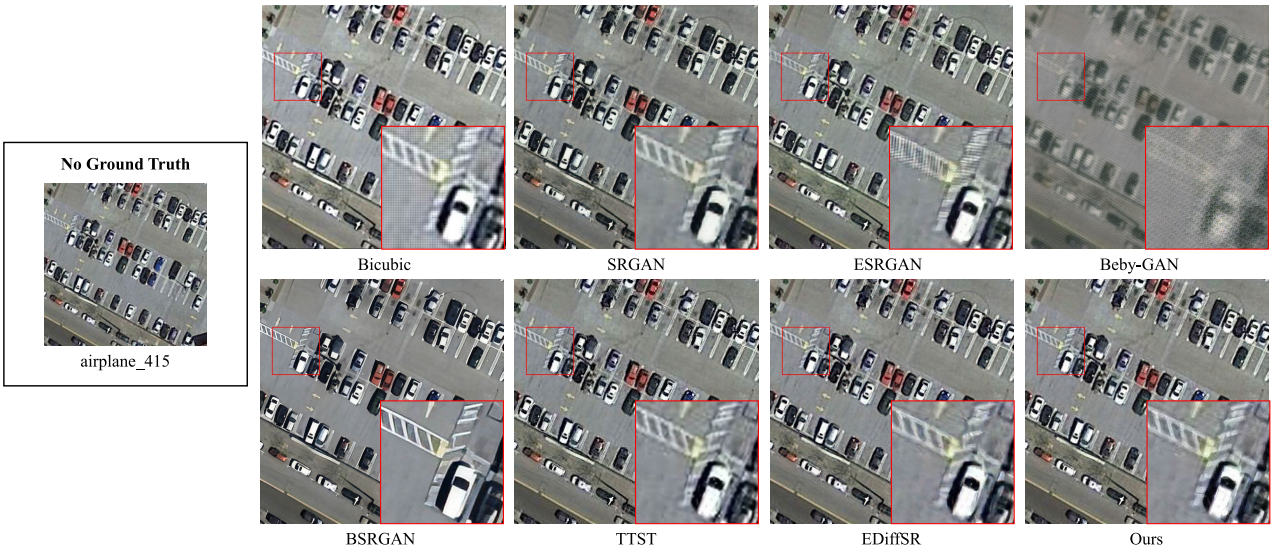

Fig. 9. 4 × visual comparison with the SOTA SR model on the NWPU-RESISC45 test set.

enabling DRGAN to recover details that are consistent with the genuine distribution.

Fig. 6 provides a closer examination of several illustrations from the WHU-RS19 effect dataset employing a fuzzy kernel of 0.6. It can be seen from the figure that SRGAN, ESRGAN, and EDiffSR cannot handle LR images with blurring well and even amplify the negative effects of blurring. The bicubic method is plagued by severe checkerboard grid artifacts in the reconstructed images. Our method produces crisp edges and more similar colors. In the red box, only our method can restore the full details of the closest actual situation.

In Fig. 7, we visualize the SR results when the fuzzy kernel is 1.2. As shown in the figure, ESRGAN visualization is slightly better than other methods. However, the texture of the hull section in bridge_04 is still unclear, and the details in forest_01 are too smooth. Our method not only restores the details of the image but also successfully removes the artifacts.

Fig. 8 illustrates the visual effect when the blur kernel is 1.8. The image parking_05 contains a dense distribution of multiscale vehicles, and our DRGAN can distinguish the vehicles better, while other methods suffer from the defect of multitarget fusion. In railwayStation_05, our method fully considers the library edge distribution with less distortion.

3) Real-World Comparison: We also evaluate the performance of our DRGAN on real-world remote sensing images. Table IV demonstrates the quantitative comparison between DRGAN and SOTA methods regarding NIQE and AG. From the table, we can see that our DRGAN achieved the best NIQE. It demonstrates the ability of our method to recover images that are consistent with human perception.

TABLE V

ABLATION ANALYSIS OF DIFFERENT MODULES. THE BEST VALUES ARE HIGHLIGHTED IN RED

	PSNR	SSIM	LPIPS	NIQE
RRDB	26.44	0.6792	0.4590	7.4215
OSRRDB	26.51	0.6816	0.4501	7.6622
MaxPool	24.92	0.6187	0.3467	3.8596
AvgPool	25.51	0.6327	0.339	4.6362

The quantitative results for the NWPU-RESISC45 are portrayed in Fig. 9. Beby-GAN provides considerable blurring when compared to other methodologies. BSRGAN results in an excessively sharpened effect. Conversely, our methodology recaptures high-frequency texture intricacies and minimizes blurring and artifacts.

E. Ablation Studies

1) Comparison of RRDB and OSRRDB: Our study compares the performance of RRDB and OSRRDB. According to the data in Table V, OSRRDB improves the PSNR by 0.07 compared to RRDB. In addition, OSRRDB outperformed RRDB by approximately 1.98% and 3.24% on the LPIPS and NIQE assessment metrics, respectively. These results indicate that OSRRDB performs superior in image reconstruction capability and visual quality assessment, showing its potential and effectiveness in image processing tasks.

2) Comparison of Pooling Layers: To explore the effectiveness of the average pooling layer, we compare the performance of the average pooling layer with the maximum pooling

layer in detail. As shown in Table V, the average pooling layer improves by 0.59 in terms of the PSNR and 2.26% in terms of SSIM compared to the maximum pooling layer. In addition, the LPIPS score was also 2.22% higher than the maximum pooling layer, while the NIQE score improved by 20.12%. These results clearly show the significant advantages of average pooling layers in image processing. It not only maintains image quality better but also performs better in terms of detail retention and visual perception.

V. Conclusion

In this article, we design a DRGAN SR reconstruction model, aiming to generate SR results that conform to human eye perception. To recover remote sensing image details to a greater extent, we design an OSRRDBNet that deeply extracts and reconstructs input features to enhance the detailed information and image quality. Extensive testing on remote sensing datasets subjected to various degradations shows that our DRGAN outperforms current SOTA methods.

Nevertheless, our DRGAN model does have certain drawbacks. First, the training process of GAN models is computationally expensive, hindering real-time performance. Besides, DRGAN may still suffer from the problem of detail loss in practical applications, especially in complex multidegradation scenarios. Therefore, these challenges should be gradually addressed in the future by optimizing the algorithm and improving the training strategy to enhance the performance and usefulness of the model.

References

[1] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.

[2] K. Chen et al., "Continuous remote sensing image super-resolution based on context interaction in implicit function space," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702216.

[3] Y. Wang, W. Liu, W. Sun, X. Meng, G. Yang, and K. Ren, "A progressive feature enhancement deep network for large-scale remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5619413.

[4] R. Neyns and F. Canters, "Mapping of urban vegetation with high-resolution remote sensing: A review," *Remote Sens.*, vol. 14, no. 4, p. 1031, Feb. 2022.

[5] J. Li et al., "Deep learning in multimodal remote sensing data fusion: A comprehensive review," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 112, Aug. 2022, Art. no. 102926.

[6] Y. Wu, Z. Li, B. Zhao, Y. Song, and B. Zhang, "Transfer learning of spatial features from high-resolution RGB images for large-scale and robust hyperspectral remote sensing target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5505732.

[7] L. Sun, J. Pan, and J. Tang, "ShuffleMixer: An efficient ConvNet for image super-resolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 17314–17326.

[8] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, and L. Zhang, "TTST: A top-k token selective transformer for remote sensing image super-resolution," *IEEE Trans. Image Process.*, vol. 33, pp. 738–752, 2024.

[9] M. Ibrahim, R. Benavente, D. Ponsa, and F. Lumbreras, "SWViT-RRDB: Shifted window vision transformer integrating residual in residual dense block for remote sensing super-resolution," in *Proc. 19th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, Jan. 2024, pp. 575–582.

[10] Y. Xiao, Q. Yuan, K. Jiang, J. He, X. Jin, and L. Zhang, "EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5601514.

[11] S. Lei and Z. Shi, "Hybrid-scale self-similarity exploitation for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–10, 2021.

[12] G. Wu, J. Jiang, K. Jiang, and X. Liu, "Fully 1×1 convolutional network for lightweight image super-resolution," *Mach. Intell. Res.*, vol. 21, pp. 1–15, 2024.

[13] H. Li et al., "SRDiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, Mar. 2022.

[14] B. Xia et al., "DiffIR: Efficient diffusion model for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 13095–13105.

[15] J. He, Q. Yuan, J. Li, Y. Xiao, X. Liu, and Y. Zou, "DsTer: A dense spectral transformer for remote sensing spectral super-resolution," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 109, May 2022, Art. no. 102773.

[16] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.

[17] X. Wang, L. Sun, A. Chehri, and Y. Song, "A review of GAN-based super-resolution reconstruction for optical remote sensing images," *Remote Sens.*, vol. 15, no. 20, p. 5062, Oct. 2023.

[18] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[19] X. Wang et al., "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Jan. 2019, pp. 63–79.

[20] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 4791–4800.

[21] C. Dong, C. 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. 2015.

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

[23] J. Yu, Y. Fan, and T. S. Huang, "Wide activation for efficient image and video super-resolution," in *Proc. 30th Brit. Mach. Vis. Conf. (BMVC)*, Jan. 2019, p. 189.

[24] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 286–301.

[25] J. Liang, H. Zeng, and L. Zhang, "Efficient and degradation-adaptive network for real-world image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 574–591.

[26] H. Zhang, C. Zhang, F. Xie, and Z. Jiang, "A closed-loop network for single infrared remote sensing image super-resolution in real world," *Remote Sens.*, vol. 15, no. 4, p. 882, Feb. 2023.

[27] H. Zhang, P. Wang, and Z. Jiang, "Nonpairwise-trained cycle convolutional neural network for single remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4250–4261, May 2021.

[28] H. Zhou et al., "Learning correction filter via degradation-adaptive regression for blind single image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 12331–12341.

[29] S. Lei, Z. Shi, and W. Mo, "Transformer-based multistage enhancement for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2021.

[30] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: Image restoration using Swin transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1833–1844.

[31] Z. Luo, F. K. Gustafsson, Z. Zhao, J. Sjölund, and T. B. Schön, "Image restoration with mean-reverting stochastic differential equations," 2023, *arXiv:2301.11699*.

[32] Z. Yue, J. Wang, and C. C. Loy, "ResShift: Efficient diffusion model for image super-resolution by residual shifting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 13294–13307.

[33] W. Li, K. Zhou, L. Qi, L. Lu, and J. Lu, "Best-buddy GANs for highly detailed image super-resolution," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 2, pp. 1412–1420.

[34] R. Dong, L. Zhang, and H. Fu, "RRSGAN: Reference-based super-resolution for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5601117.

[35] B. Li et al., "SeD: Semantic-aware discriminator for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 25784–25795.
[36] M. Y. Yang, W. Liao, X. Li, Y. Cao, and B. Rosenhahn, "Vehicle detection in aerial images," *Photogramm. Eng. Remote Sens.*, vol. 85, no. 4, pp. 297–304, Apr. 2019.
[37] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.
[38] W. Huang, Q. Wang, and X. Li, "Feature sparsity in convolutional neural networks for scene classification of remote sensing image," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 3017–3020.
[39] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.
[40] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004.
[41] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[42] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6626–6637.
[43] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.

Yongchao Song was born in Weihai, Shandong, China, in 1990. He received the B.S. and Ph.D. degrees from the School of Electronic and Control Engineering, Chang'an University, Xi'an, China, in 2015 and 2020, respectively.

He is currently an Associate Professor with the School of Computer and Control Engineering, Yantai University, Yantai, China. His research interests include remote sensing information processing and deep learning.

Lijun Sun was born in Yantai, Shandong, China, in 2001. She received the B.S. degree from the School of Computer and Control Engineering, Yantai University, Yantai, China, in 2023, where she is currently pursuing the master's degree.

Her research interests include deep learning, remote sensing image super-resolution, and artificial intelligence.

Jiping Bi was born in Weifang, Shandong, China, in 2001. He received the B.S. degree from the School of Computer and Control Engineering, Yantai University, Yantai, China, in 2023, where he is currently pursuing the master's degree.

His research interests include image processing, traffic target detection, and automatic control.

Siwen Quan received the B.S. degree from Chang'an University, Xi'an, China, in 2015, and the Ph.D. degree from Huazhong University of Science and Technology, Wuhan, China, in 2019.

She is an Associate Professor with the School of Electronics and Control Engineering, Chang'an University. Her research interests include local geometric shape description, 3-D object recognition, and image fusion.

Xuan Wang (Senior Member, IEEE) was as born in Weihai, Shandong, China, in 1991. She received the B.S. and Ph.D. degrees from Traffic Information Engineering and Control, Chang'an University, Xi'an, China, in 2013 and 2018, respectively.

She is currently an Associate Professor with the School of Computer Science and Control Engineering, Yantai University, Yantai, China. Her research interests include intelligent traffic control, artificial intelligence, and computer vision.