electronics

MDPI

Article

ESTUGAN: Enhanced Swin Transformer with U-Net Discriminator for Remote Sensing Image Super-Resolution

Chunhe Yu, Lingyue Hong *, Tianpeng Pan, Yufeng Li and Tingting Li

School of Electronical and Information Engineering, Shenyang Aerospace University, Shenyang 110136, China; 20052801@sau.edu.cn (C.Y.); pantianpeng@stu.sau.edu.cn (T.P.); liyufeng@sau.edu.cn (Y.L.); litingting@stu.sau.edu.cn (T.L.)

* Correspondence: honglingyue@stu.sau.edu.cn

Abstract: Remote sensing image super-resolution (SR) is a practical research topic with broad applications. However, the mainstream algorithms for this task suffer from limitations. CNN-based algorithms face difficulties in modeling long-term dependencies, while generative adversarial networks (GANs) are prone to producing artifacts, making it difficult to reconstruct high-quality, detailed images. To address these challenges, we propose ESTUGAN for remote sensing image SR. On the one hand, ESTUGAN adopts the Swin Transformer as the network backbone and upgrades it to fully mobilize input information for global interaction, achieving impressive performance with fewer parameters. On the other hand, we employ a U-Net discriminator with the region-aware learning strategy for assisted supervision. The U-shaped design enables us to obtain structural information at each hierarchy and provides dense pixel-by-pixel feedback on the predicted images. Combined with the region-aware learning strategy, our U-Net discriminator can perform adversarial learning only for texture-rich regions, effectively suppressing artifacts. To achieve flexible supervision for the estimation, we employ the Best-buddy loss. And we also add the Back-projection loss as a constraint for the faithful reconstruction of the high-resolution image distribution. Extensive experiments demonstrate the superior perceptual quality and reliability of our proposed ESTUGAN in reconstructing remote sensing images.

Keywords: Swin Transformer; U-Net discriminator; remote sensing image; super-resolution; generative adversarial network

Citation: Yu, C.; Hong, L.; Pan, T.; Li, Y.; Li, T. ESTUGAN: Enhanced Swin Transformer with U-Net Discriminator for Remote Sensing Image Super-Resolution. *Electronics* **2023**, *12*, 4235. https://doi.org/10.3390/electronics12204235

Academic Editor: Byung Cheol Song

Received: 28 August 2023
Revised: 30 September 2023
Accepted: 11 October 2023
Published: 13 October 2023

1. Introduction

The rapid development of modern aerospace technology has put remote sensing imagery into wider use in the remote sensing field. Remote sensing images are essential for applications such as target detection and tracking. However, obtaining high-resolution (HR) remote sensing images can be challenging due to technical limitations and cost constraints. Image super-resolution (SR) is a promising option and a heated technology in recent years that provides critical research significance. In recent years, deep learning-based methods for single image super-resolution (SISR) have made remarkable achievements. Since the proposal of the SRCNN [1] by Dong et al. in 2014, CNN-based methods have significantly advanced the field of SR. Scholars have continuously improved network architecture and proposed elaborate structures [2–4], such as residual learning, dense connectivity, Laplace pyramid, and so on. RCAN [5] has achieved another pinnacle of peak signal-to-noise ratio (PSNR) by adding the channel attention module to the CNN-based architecture. However, CNN-based methods face an unavoidable obstacle when it comes to SR. Due to the design of the convolutional layer, convolution kernels interact with the image in a content-independent process. It is illogical to use the same convolutional kernel to reconstruct different areas of the image. The transformer architecture [6–9] stands out in this case, employing the self-attention mechanism for global interaction and achieving significant

performance in several visual tasks. However, due to the quadratic complexity of processing images, transformer-based models tend to generate a large number of parameters and are computationally intensive. The Swin Transformer [10] was created to combine the advantages of transformer- and CNN-based models, not only establishing long-term dependencies between images, but also processing large-sized images through a local attention mechanism. SwinIR [11] firstly applies the Swin Transformer to the field of SISR; it achieves the optimal PSNR with fewer parameters and is an enormous prospect. HAT [12] activates more input signals by concatenating the channel attention mechanism in the Swin Transformer layer and proposes the overlapping cross-window attention mechanism to optimize cross-window information interaction.

While the methods mentioned above have achieved high PSNR scores, they can produce ambiguous results. This is because they often use MSE or MAE for the one-to-one supervision of a single low-resolution (LR) image corresponding to a single high-resolution (HR) image, which can lead to pixel averaging and overly smooth and blurred outcomes. Remote sensing images are mainly used in the fields of object detection as well as geologic analysis, and we believe that the over-smoothed and blurred results generated by these networks will have a negative impact on some of the categories. To obtain more realistic images, researchers have employed Generative Adversarial Networks (GANs) to recover images with rich texture details [13–16]. Although these methods have made considerable progress, further research is necessary due to their difficulty in training and tendency to produce artifacts. An alternative approach proposed by [17] is the Best-buddy loss, which breaks the strict mapping between LR and HR set by MSE or MAE. This approach allows multiple patches close to ground truth to supervise SR, reducing the difficulty of network training while improving the perceptual quality of reconstructed images.

The learning-based approaches mentioned above offer a new development direction for the remote sensing image SR task. LGCNet [18] is the first CNN-based SR model for remote sensing images that outperforms traditional methods and verifies the effectiveness of deep learning methods. Jiang et al. [19] propose an edge enhancement network based on a GAN to enhance the edge by learning noise masks. Some algorithms [20–26] have achieved considerable success by adding elaborate structural designs or various attention mechanisms to CNN. Currently, learning-based methods in remote sensing image SR are developing rapidly and have achieved remarkable progress, but the challenges are still significant.

The selection of a reconstruction network better suited to the characteristics of remote sensing images is a challenging problem, because remote sensing images are characterized by a large spatial span, complex texture structure, and few pixels covered by objects, which undoubtedly produce further difficulties to reconstruction tasks [27]. To faithfully restore high-resolution images, we adopt the Swin Transformer as the backbone, which can realize long-term dependency modeling with shift windows and exploit the internal self-similarity within remote sensing images. Specifically, we adopt the Residual Hybrid Attention Group (RHAG) proposed by HAT [12] and refine its network design to obtain significant performance with fewer parameters, which is named the Enhanced Swin Transformer Network (ESTN).

However, simply utilizing a more powerful reconstruction network will not completely achieve satisfactory results in the remote sensing image SR task. This is because objects in remote sensing images cover fewer pixels, and a ship may be represented by only several pixels. Employing PSNR-based methods is vulnerable to blurred results, while GANs offer a decent solution. In addition, remote sensing images contain more diverse texture features and different regions with distinct texture differences [27]. We discovered that regions with different texture complexity in remote sensing images should not adopt the same supervision strategy. Adversarial learning should be performed for texture-rich regions to facilitate the reconstruction of fine details. However, for the smooth region, the PSNR-based method is sufficient to recover satisfactory results. Instead, feeding such regions into the discriminator may lead to uncomfortable artifacts. Existing methods do not take this concern into account. To resolve the above problems, we propose the U-Net

discriminator with the region-aware learning strategy. On the one hand, the U-shaped network design allows the discriminator to fully integrate the structural information at each hierarchy level and finally obtain pixel-by-pixel feedback. On the other hand, it can divide the areas according to texture complexity, and only the detailed regions are fed into the discriminator, forcing the discriminator to focus on distinguishing complex areas and greatly suppressing artifacts. Accordingly, our discriminator can effectively assist the ESTN in predicting realistic and highly detailed images.

To further improve the perceptual quality, we also introduce the Best-buddy (BB) loss [17] and Back-projection (BP) loss to break the rigid mapping from the LR space to the HR space. This reduces the training difficulty and contributes to the recovery of realistic texture details.

Overall, the main contributions of our work are as follows:

(1) We propose a promising framework, ESTUGAN, which adopts the Enhanced Swin Transformer as the generator backbone and a U-Net discriminator. The Enhanced Swin Transformer is capable of mobilizing more input information to model local content, benefiting from united channel attention and self-attention. In addition, it employs an overlapping cross-attention mechanism to further aggregate cross-window information with stronger representational capabilities. Extensive experiments demonstrate that our proposed network outperforms other methods when targeting remote sensing image SR.
(2) We propose a U-Net discriminator with the region-aware learning strategy to reconstruct highly detailed remote sensing images. The region-aware learning strategy can effectively suppress artifacts by masking flat regions and feeding only texture-rich regions to the discriminator for adversarial training. Moreover, the U-shaped network is designed with jumping connections that allows for the connection of shallow detailed content with deep semantic information, providing intensive feedback for each pixel's authenticity.
(3) The BB loss and BP loss are employed to further enhance the visual quality of the image. Multiple supervised signals that are similar to the ground truth are utilized to flexibly guide the image reconstruction; this reduces the training difficulty and helps to generate high-frequency information.

2. Related Works

The following contents list some aspects of the previously proposed methodology related to our proposed ESTUGAN:

2.1. Swin Transformer

The Swin Transformer [10] is a universal backbone for vision tasks and represents one of the first hierarchical vision transformers. Due to its excellent performance and parallelization accessibility, it has become the state-of-the-art technology for various vision tasks such as target detection and image segmentation. The core idea of the Swin Transformer is to compute self-attention within a non-overlapping movable window, which makes the model computation linear with respect to the feature map resolution, and greatly compresses the cost of self-attention. SwinIR introduces the Swin Transformer to image SR for the first time, further refreshing the state of the art of SR tasks. However, there is still substantial room for improvement in the Swin Transformer. The window attention mechanism [28–30] has limitations, and the exchange of information across windows and the shallow message mobilization both require further optimization.

2.2. Generative Adversarial Network

Nowadays, GANs have been widely explored and have achieved remarkable achievements in various image processing domains such as style migration, super resolution, image complementation, and denoising tasks [31–33]. This approach is mainly inspired by the idea of competition in game theory, which is applied to deep learning by construct-

ing two deep learning models: a generative network G (generator) and a discriminator network D (discriminator). The two models are then continuously played against each other to make G generate realistic images, while D has the powerful ability to determine the image authenticity. To reconstruct images with high perceptual quality, SRGAN introduces a discriminator that guides the generator to recover the fine texture information by adversarial loss. ESRGAN [15] proposes Residual-in-Residual Dense Blocks (RRDB) to build the network and invokes the relativistic GAN to make discriminators predict relative truthfulness, winning first prize in the PIRM 2018-SR Challenge. These approaches have been widely adopted as the mainstream of perception-based image SR algorithms.

2.3. Loss Function on Deep Learning

It is obvious that SISR is inherently an ill-posed problem, where a LR image often corresponds to multiple HR images. Proper guidance of the model to find the region in the latent space closest to the real HR image is the key to the SR problem. Therefore, a suitable loss function becomes particularly relevant. In existing studies, most algorithms adopt MAE/MSE loss to make the SR image approximate to the ground truth pixel by pixel. This pixel-level loss is beneficial to upgrade the PSNR but is detrimental to the reconstruction of texture details [34]. To solve this problem, perceptual loss [35] is proposed to compute the similarity of deep features to enhance the perceptual quality. Fuoli et al. [36] propose Fourier spatial loss to facilitate the recovery of lost high frequency information. Benefiting from perceptual loss and adversarial loss, SRGAN [13,14] and ESRGAN [15] recover photo-realistic outcomes, but they face the possibility of annoying artifacts. Liang et al. introduce the Local Discriminant Learning (LDL) strategy [37] that explicitly penalizes artifacts without sacrificing real details, alleviating the artifact problem partly. Li et al. suggest the Best-buddy loss [17] to address the above problems. The estimated patches are enabled to seek optimal supervision dynamically during training, contributing to the production of more reasonable details.

2.4. Deep Learning Based SISR for Remote Sensing Images

In recent years, deep learning based SISR has become mainstream due to the powerful extraction capabilities of deep neural networks. And these approaches also lead to the development and advancement of remote sensing image SR algorithms. The CNN-based SISR was widely adopted by scholars in the early days; they retrained the network on remote sensing images and designed elaborate network architectures for feature extraction. LGCNet [18] learns hierarchical representations of remote sensing images by constructing a "multifork" structure. DDRN [38] proposes ultra-dense residual blocks to construct a simple but effective recursive network. Similarly, many refined structural designs have been applied to the network with impressive achievements. However, the convolutional kernel interacts with the image in a content-independent manner, which limits the reconstruction of texture details. Some works enhance the expressive power of the model by adding various attentional mechanisms, such as MHAN [39] and SMSR [40]. But these approaches tend to be computationally intensive and still have long-term dependency modeling difficulties. In addition, the above method adopts the learning strategy which maximizes the PSNR and encourages the model to find the pixel mean, leading to blurred results. Regarding this topic, several related works have made promising progress. On the one hand, adversarial learning strategies have been employed by some works, such as SRGAN and ESRGAN, in order to reconstruct photo-realistic images. MA-GAN [27] and SRAGAN [41] combined a GAN with attention mechanisms to upgrade the visual quality of remote sensing images. On the other hand, some loss functions [35–37] have been proposed to motivate the generation of high-frequency content. However, these solutions are still not perfect, since problems remain, like the difficulty of GAN training and the potential for artifacts. Our work is based on a GAN, which employs the Swin Transformer as the generator for long term dependency modeling, and a U-Net discriminator with

the region-aware strategy to facilitate high-frequency detail generation while suppressing artifacts to a certain extent.

2.5. Image Super Resolution Quality Assessment

SR image quality assessment is an effective way to evaluate and compare SR methods, which is an important guide for model optimization and parameter selection. Subjective human assessment represents a highly reliable evaluation approach, but it tends to be time-consuming and laborious. The PSNR [42] is the most popular metric to assess the reconstruction performance by calculating only the purely mathematical difference of pixels. Wang et al. [43] simulate the human visual system and propose an evaluation scheme based on structural similarity. However, these two options sometimes differ from the human eye's perceptual quality, leading to ambiguous predictions. In order to maintain better consistency with subjective quality evaluations, a comparison of the feature similarity between images is employed by Zhang et al. [44] to estimate the distance from the prediction to the ground truth. The SFSN model [45] aims to find a balance between structural fidelity and statistical naturalness. Then, SRIF [46] is proposed to merge deterministic fidelity and statistical fidelity into a single prediction. Thanks to the development of deep learning, Ref. [47] extracts deep features to appraise the Learned Perceptual Image Patch Similarity (LPIPS) between two images, which is more in line with the human perceptual situation. DeepSRQ [48] with deep two-stream convolutional networks provides a satisfactory solution to the problem of no-reference evaluation.

3. Methods

In this section, we first present a brief overview on the workflow of our algorithm, and then we give a detailed description for the generator, the U-Net discriminator with the region-aware learning strategy, and the loss function employed by ESTUGAN, respectively.

3.1. Overview of ESTUGAN

For recovering images with superior perceptual quality, we designed the ESTUGAN based on a GAN, which consists of the ESTN as the generator, and the U-Net discriminator. The principal framework is shown in Figure 1. Given an LR image $I_{LR} \in R^{H \times W \times C}$, an SR image $I_{SR} \in R^{rH \times rW \times C}$ (r is the scale factor) can be obtained by the generator, denoted as

$$I_{SR} = G(I_{LR}) \tag{1}$$

where $G(\cdot)$ denotes the generator. Subsequently, unlike the approach of [14], which feeds I_{SR} directly to the discriminator, in our approach, I_{SR} is sent to the region-aware adversarial learning stage, where we feed only regions with rich texture details to the U-Net discriminator for authenticity judgments by regional division processing. Finally, the discriminator outputs the real probability map and feeds it back to the generator, prompting the generation of real abundant details. In a GAN, the generator is urged to deceive the discriminator by creating realistic fake HR images, while the discriminator is trained to be powerful in discriminating authenticity, and both of them compete against each other to make the SR image distribution gradually approximate the real image distribution.

3.2. The Architecture of the Generator

As shown in Figure 2, we keep the high-performance architecture design of SwinIR [11], and the whole generator is composed of three modules: shallow feature extraction, deep feature extraction, and image reconstruction.

In the shallow feature extraction module, we employ a separate convolutional layer to map the input image to a high-dimensional space. It helps the visual representation to be learned better and optimized stably. The extracted shallow features can be expressed as

$$F_0 = H_{SF}(I_{LR}) \tag{2}$$

where $H_{SF}(\cdot)$ denotes the shallow feature extraction.

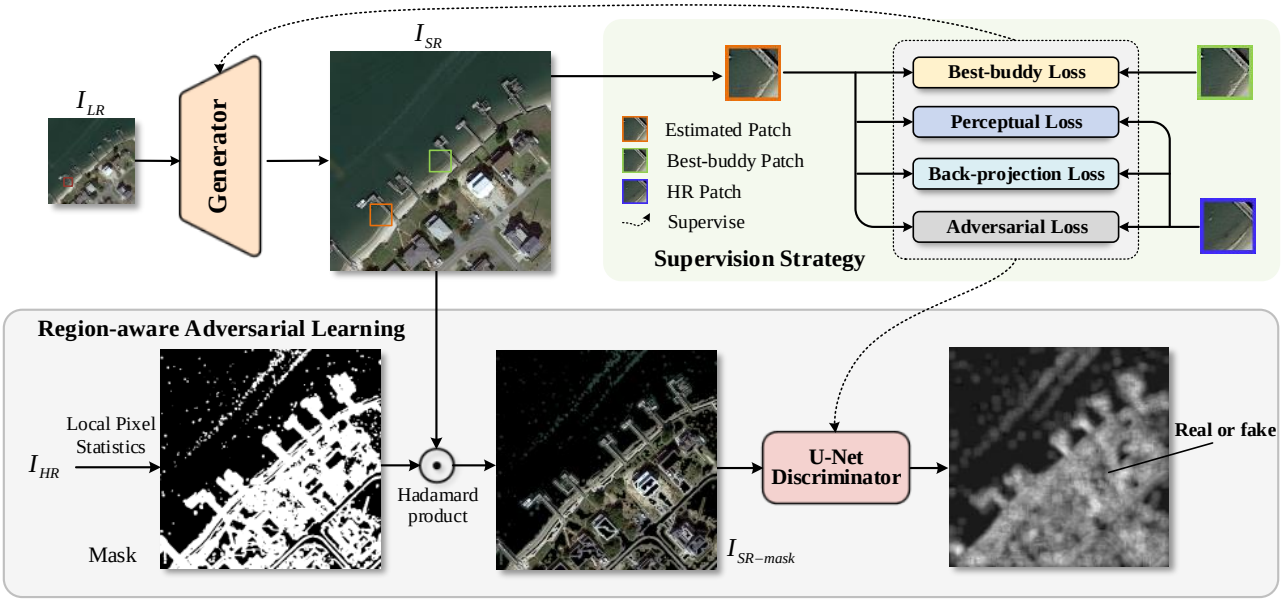
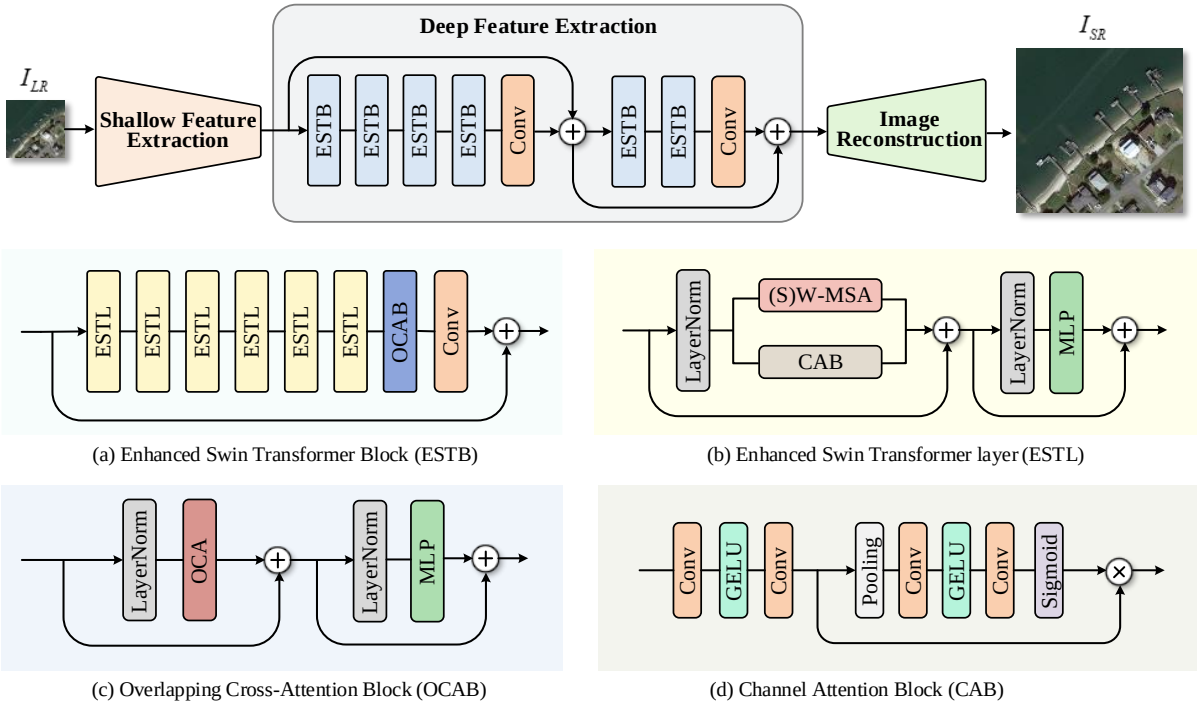


Figure 1. The overview of our proposed ESTUGAN. The proposed U-Net discriminator with region-aware learning strategy focuses on adversarial learning in texture-rich regions and outputs a map for the true situation of each pixel. We use Best-buddy loss, Back-projection loss, perceptual loss, and adversarial loss to supervise the generator, and adversarial loss to guide the optimization for the discriminator.



(a) Enhanced Swin Transformer Block (ESTB)

(b) Enhanced Swin Transformer layer (ESTL)

(c) Overlapping Cross-Attention Block (OCAB)

(d) Channel Attention Block (CAB)

Figure 2. The framework of our proposed ESTN generator, which consists of three modules in total, including the deep feature extraction module, shallow feature extraction module, and image reconstruction module.

In the deep feature extraction module, we adopt a new basic block inspired by HAT [12], called Residual Hybrid Attention Group. And we rename it to Enhanced Swin Transformer Block (ESTB) for the convenience of description, the architecture of which is shown in Figure 2a. It integrates the channel attention mechanism and the overlapping

cross-attention block (OCAB), which achieves an effective aggregation for cross-window information. In addition, we insert a second residual mechanism after the convolution kernel behind the fourth ESTB. Although the residual block [49] can increase the perceptual field, we find that in low-level reconstruction tasks, such as image SR, excessively long residual connections will on the contrary weaken the generation quality of the reconstructed images, because overly abstract high-dimensional features can make network learning more difficult and cause degradation in the performance of the generation network [50]. To further demonstrate the effect of the number of residual blocks and the number of connection dimensions on the network performance, we set up three different networks in the ablation study section to demonstrate the superior performance of our network. The processes can be formulated as follows:

$$F_{DF} = H^{i}_{DF}(F_0) + F_0 \tag{3}$$

$$F'_{DF} = H^{i}_{DF}(F_{DF}) + F_{DF} \tag{4}$$

where $H^{i}_{DF}(\cdot)$ denotes the deep feature extraction module, containing i ESTB blocks and a 3×3 convolutional layer. In this paper, i is set to 4 in Equation (3), and in Equation (4), i is set to 2.

In the image reconstruction module, we use jump connections to aggregate deep features and shallow features and reconstruct high-resolution images with the pixel-shuffle method [51]. It can be expressed as

$$I_{SR} = H_{Rec}\left(F'_{DF}\right) \tag{5}$$

where $H_{Rec}(\cdot)$ indicates reconstruction module.

3.3. U-Net Discriminator with Region-Aware Learning Strategy

As for the discriminator, inspired by [52,53], we adopt the U-Net discriminator, which essentially consists of an encoder and a decoder to be connected, as shown in Figure 3. The encoder continuously downsamples the I_{SR} in order to obtain the global information, and finally reacts to the overall image reality. While the decoder is dedicated to the local information authenticity judgment, it keeps performing progressive upsampling operations to output the per-pixel reality with the same resolution as I_{SR}. In addition, skip connections are applied to facilitate the information communication between the two networks, further promoting the detailed recovery. Such a structural design forces the discriminator to focus on the structural and semantic message differences between fake and genuine samples, pursuing the accuracy of the global context and local information of the reconstruction outcome.

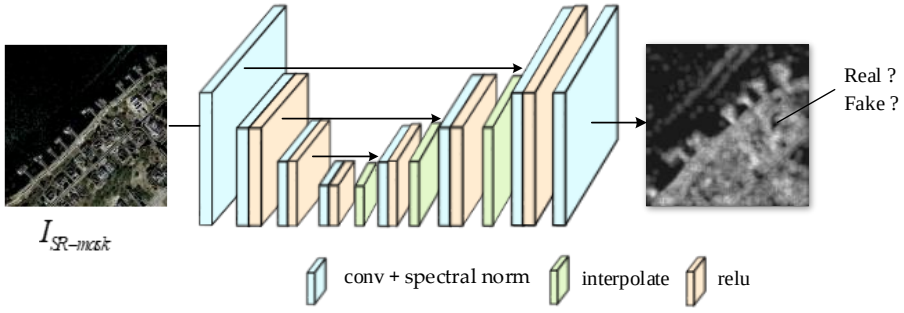


Figure 3. The framework of our proposed U-Net discriminator.

For addressing the artifacts of the GAN-based methods [17], the region aware strategy is appended within the U-Net discriminator, as shown in Figure 1. The smooth regions and texture-rich regions of I_{SR} are separated by the statistical local pixel distribution of I_{HR}, and only texture-rich regions are fed into the U-Net discriminator for adversarial learning. This not only avoids the generation of artifacts in smooth areas, but also permits

the discriminator to focus on regions where fine realistic details are required to be recovered, assisting the reconstruction of perceptually realistic images. Regarding the specific region-aware learning strategy, we first perform the unfold operation with kernel size k on I_{HR} to obtain $rH \times rW$ patches $Q_{i,j}$ with size k^2. The standard deviation $std\left(Q_{i,j}\right)$ is then calculated for each patch, and the final binary feature map $M_{i,j}$ is obtained by comparison with the pre-set threshold, which is denoted as

$$M_{i,j} = \begin{cases} 0, & std\left(Q_{i,j}\right) \leq \theta \\ 1, & std\left(Q_{i,j}\right) > \theta \end{cases} \tag{6}$$

where i and j denote the specific locations of patches, and the pixel values are set to 0 for flat regions and 1 for texture-rich regions in the map. Finally, I_{SR_mask} is obtained by multiplying $M_{i,j}$ with I_{SR}.

In addition, we also introduce the spectral normalization regularization [54] to further secure the stability of training and suppress artifacts.

3.4. Loss Function

3.4.1. Best-Buddy Loss

Since a single LR image can correspond to multiple HR images, SISR is intrinsically an indeterminate problem. For a given HR-LR pair, the commonly adopted MSE/MAE loss tends to perform a one-to-one rigid mapping, as shown in the blue diagram of Figure 4. This overlooks the intrinsic uncertainty of SISR, resulting in reconstructed images lacking high-frequency information. In order to overcome the limitation caused by the supervision of I_{SR} from a single I_{HR}, we refer to [55–59] and adopt the BB loss. It allows diverse supervised patches $p^i_{hr^*}$ to positively steer the predicted patches p_{sr} and achieves the multiplicity of supervision, as shown in the yellow diagram of Figure 4. For $p^i_{hr^*}$, it should be as close as possible to both the predicted patches p_{sr} and the patch p_{hr} of I_{HR}, which can be expressed as

$$p^i_{hr^*} = \underset{p \in B}{\operatorname{argmin}} \left\| p - p^i_{hr} \right\|_2^2 + \left\| p - p^i_{sr} \right\|_2^2 \tag{7}$$

where $\|\cdot\|_2$ expresses L_2 loss, B denotes the supervised candidate database [17] of this image, which is obtained from the three-level image pyramid expansion achieved by the bicubic downsample operation, and i denotes the number of iterations. Then, the BB loss of this patch can be expressed as

$$L_{BB}\left(p^i_{sr}, p^i_{hr^*}\right) = \left\| p^i_{sr} - p^i_{hr^*} \right\|_1 \tag{8}$$

where $\|\cdot\|_1$ denotes L_1 loss.

3.4.2. Adversarial Loss

Adversarial loss is employed to facilitate perceptually realistic image generation, and the adversarial loss of the generator and discriminator are respectively denoted as

$$L_{adv\_G} = L_{BCE}(D(I_{SR}), U_{real}) \tag{9}$$

$$L_{adv\_D} = L_{BCE}(D(I_{HR}), U_{real}) + L_{BCE}\left(D(I_{SR}), U_{fake}\right) \tag{10}$$

where $L_{BCE}(\cdot)$ denotes binary cross entropy loss, $D(\cdot)$ denotes the output of the discriminator, which is a tensor of shape $rH \times rW \times 1$, U_{real} and U_{fake} are tensors with the same shape as $D(\cdot)$, where all the values of U_{real} are 1 for real labels and all the values of U_{fake} are 0 for fake labels.

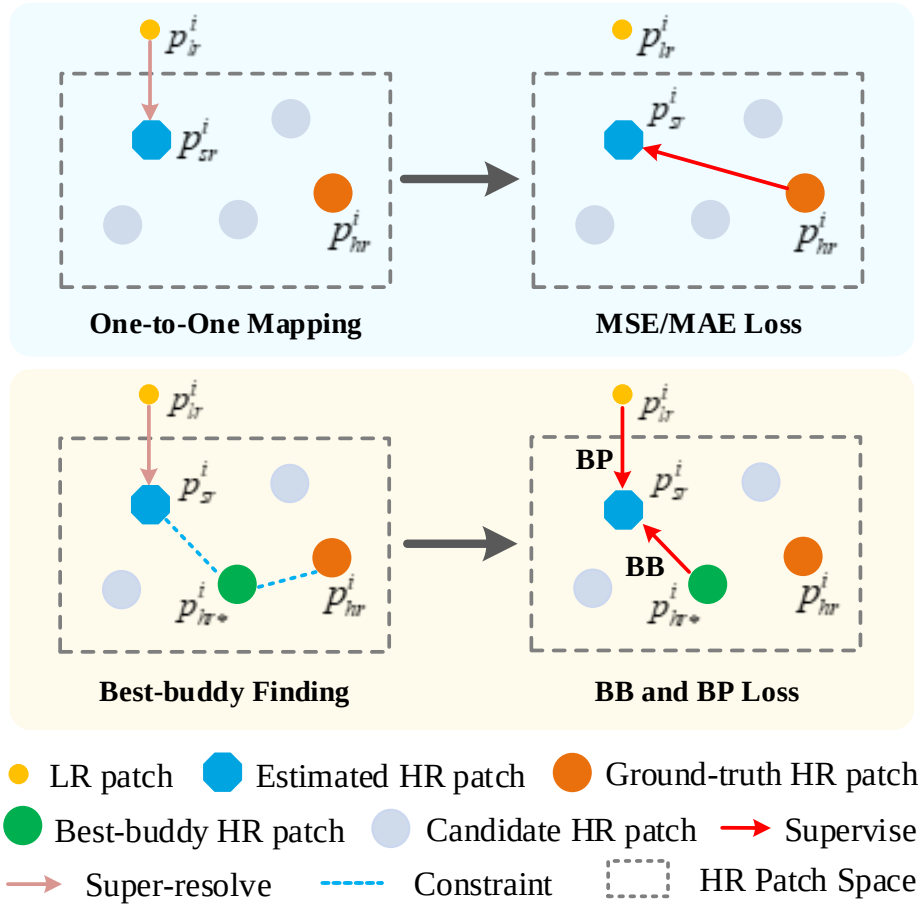


Figure 4. Our Best-buddy (BB) loss combined with Back-projection (BP) loss for supervision compared to MSE/MAE loss. Specifically, the blue plot represents the MAE/MSE loss, and the yellow plot represents the BB loss and BP loss we adopted. p^i_{lr} , p^i_{hr}, p^i_{sr}, and $p^i_{hr^*}$ denote the LR patch, HR patch (ground truth), predicted HR patch, and Best-buddy HR patch in the current iteration, respectively.

3.4.3. Perceptual Loss

The perceptual loss is calculated utilizing the three layers, $conv_{3-4}$, $conv_{4-4}$, and $conv_{5-4}$, of the feature maps in the pre-trained VGG19 network, which can be expressed as

$$L_p = \sum_{i=3}^{5} \alpha_i \| conv_{i-4}(I_{SR}) - conv_{i-4}(I_{HR}) \|_1 \tag{11}$$

where α_i denotes the weight occupied by each layer, and $\alpha_3 = 1/8$, $\alpha_4 = 1/4$, and $\alpha_5 = 1/2$, respectively.

3.4.4. Back-Projection Loss

The adoption of BP loss forces the LR image obtained by downsampling I_{SR} with r times to match I_{LR}, achieving further supervision for I_{SR} in the low-resolution image space, which can be denoted as

$$L_{BP} = \| bi(I_{SR}, r) - I_{LR} \|_1 \tag{12}$$

where $bi(\cdot, r)$ denotes the bicubic downsampling operation with a scale factor r.

Thus, the overall generator loss can be expressed as

$$L_G = \mu_1 L_{BB} + \mu_2 L_{adv\_G} + \mu_3 L_p + \mu_4 L_{BP} \tag{13}$$

4. Experiments and Analysis

4.1. Datasets in Experiments

To validate the effectiveness of our proposed method, we selected four public remote sensing datasets, including the NWPU-RESISC45 dataset [60], the UCMerced dataset [61], the RSCNN7 dataset [62], and the DOTA dataset [63]. These datasets all consist of numerous RGB images and are extensively adopted in the remote sensing image SR field.

4.1.1. NWPU-RESISC45 Dataset

This dataset encompasses 45 classes of remote sensing images with high inter-class similarity and intra-class diversity. It contains a total of 31,500 images with a resolution of 256 × 256 pixels. We randomly selected 10 images in each category as the testing set for our experiments and used the rest as the training set. Some of the training set images are shown in Figure 5.

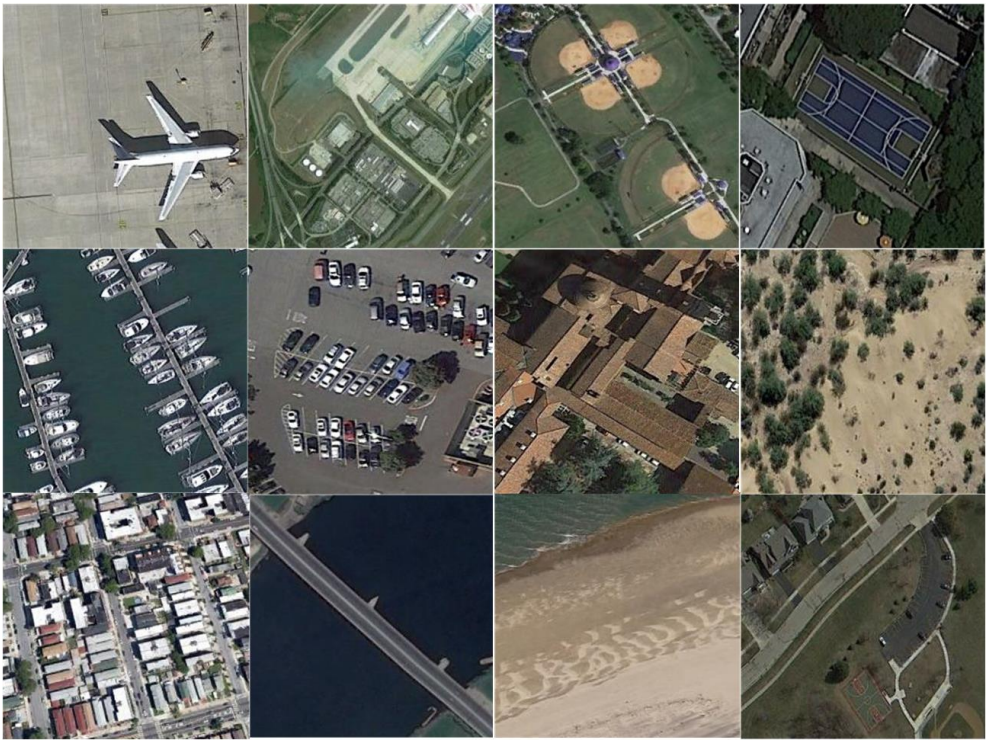


Figure 5. Partial images of the training set in the NWPU-RESISC45 dataset.

4.1.2. UCMerced Dataset

The UCMerced dataset is widely adopted for remote sensing image visual processing tasks, consisting of 21 categories, with 100 images per category. The images were captured by the remote sensing satellites of the University of California, Merced, and have a resolution of 256 × 256 pixels, covering various scenes, such as urban areas, forests, and farmlands. We randomly selected 10 images in each category as the testing set for our experiments, which can test the effectiveness of our approach and its robustness after training on the NWPU-RESISC45 dataset.

4.1.3. RSCNN7 Dataset

The RSCNN7 dataset consists of seven categories covering 2800 images and each image has 400 × 400 pixels. This dataset is sampled at different scales and takes into account weather variability and seasonal changes.

4.1.4. DOTA Dataset

The DOTA dataset consists of 2806 aerial images, each with pixel sizes ranging from 800 × 800 to 4000 × 4000, containing objects in various scales, shapes, and orientations. These images are annotated for 15 common target categories, including airplanes, ships, storage tanks, baseball fields, tennis courts, basketball courts, surface runways, harbors, bridges, large vehicles, small vehicles, helicopters, roundabouts, soccer fields, and basketball courts.

4.2. Quantitative Evaluation Metrics

In this paper, we judge the various methods using three typical image quality evaluation metrics, which are the peak signal-to-noise ratio (PSNR), the structure similarity index measure (SSIM), and the learned perceptual image patch similarity (LPIPS).

4.2.1. PSNR

The PSNR [42] is a common measure of signal reconstruction quality, and it is often defined simply by the mean squared error (MSE). For two monochrome images I and K with a size of $m \times n$, their mean squared differences are defined as

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left|\left|I(i,j) - K(i,j)\right|\right|^2 \tag{14}$$

Thus, the PSNR can be expressed as

$$PSNR = 10log_{10}\left(\frac{MAX_I^2}{MSE}\right) \tag{15}$$

where MAX_I denotes the maximum pixel value in image I, and a higher PSNR value means less distortion.

4.2.2. SSIM

The SSIM [43] is also a full-reference image quality evaluation criterion, which measures image similarity in terms of brightness, contrast, and structure, respectively. It can be expressed as

$$SSIM = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_x\sigma_y + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{16}$$

where μ_x and μ_y denote the mean pixel values for the two images, respectively, σ_x and σ_y denote the standard deviation for each image, and C_1 and C_2 are constants. The SSIM value ranges from 0 to 1, and the higher the value, the less the image distortion.

4.2.3. LPIPS

The LPIPS [47] evaluates the perceptual similarity between images according to a deep learning model, which corresponds more closely to human perception than the PSNR and SSIM do [34]. The LPIPS can be expressed as

$$LPIPS(I_{HR}, I_{SR}) = \sum_l \frac{1}{n_l}\left\|\omega_l \bigodot (\phi(I_{HR})_l - \phi(I_{SR})_l)\right\|_2^2 \tag{17}$$

where $\phi(\cdot)_l$ indicates the feature map of the l-th convolutional layer, and n_l denotes the quantity of elements in $\phi(\cdot)_l$. $\bigodot$ denotes the product operation in the channel dimension, and ω_l represents a learned weight vector. A lower value of LPIPS means that the two images are more similar in human perception.

4.3. Experimental Details

Our experiment was conducted on the NVIDIA Tesla V100 GPU. The input image size was set to 48 × 48 and the batch size was eight. We employed the bicubic operation to downsample the original high-resolution image to obtain the HR-LR training pair. The channel of our ESTN was set to sixty, and the attention heads and the window size were set to six and sixteen, respectively.

Adam was set as our optimizer and $\beta_1 = 0.9$, $\beta_2 = 0.999$, the learning rate was 1×10^{-4} while the initial stage utilized preheating and cosine decay. k and θ were introduced in the method were are set to 11 and 0.025, respectively. As for the loss function, μ_1, μ_3, and μ_4 were set to 1, while μ_2 was set to 0.005 (refer to [17]).

4.4. Comparison with State-of-the-Art Methods

4.4.1. Quantitative Comparison

In our experiments, we validated the performance of our model ESTUGAN by comparing it with six deep-learning SR methods, including RCAN [5], RRDB, SwinIR [11], SRGAN [14], ESRGAN [15], and BebyGAN [17]. We selected 31050 images from the NWPU-RESISC45 dataset as the training set and 450 images as the testing set. In addition, to verify the generalizability of these models, we included 210 randomly selected images in the UCMerced dataset, 800 randomly selected images in the DOTA dataset, and all the images in the RSCNN7 dataset as additional test sets. Under the same conditions, we tested all the methods with the 4× amplification and evaluated them using the PSNR, SSIM, and LPIPS metrics.

Table 1 shows the quantitative results. It can be seen that the proposed approach achieves the most satisfactory results. In the comparison with the GAN-based methods (SRGAN, ESRGAN, and BebyGAN), ESTUGAN achieves the maximum PSNR and SSIM, and achieves the lowest LPIPS, demonstrating that it reconstructs images with optimal accuracy and perceptual quality. It is worth mentioning that ESTUGAN still maintains the best performance on three additional test sets, validating the scalability of our proposed model. In contrast, the performance of SRGAN on the DOTA dataset shows a distinct decline, reflecting the model's shortcomings in generalizability. In the comparison with CNN-based methods (RCAN, RRDB, SwinIR), the proposed method also achieves amazing results, just slightly lower than SwinIR and higher than the other compared methods. Although it is slightly lower than SwinIR in performance, the number of parameters and FLOPs of our method are only one-fourth of those of SwinIR (illustrated in Section 4.6). The proposed method greatly saves computational resources and efficiency in the SR task for remote sensing images. The ESTN also achieves quite robust results with minimal parameters when evaluated using three additional test sets.

We also compared these methods on 45 category scenarios from the NWPU-RESISC45 dataset; as shown in Table 2, ESTUGAN outperforms the comparison methods for each scenario. Among them, the PSNR of ESTUGAN, in several scenes such as aircraft, desert, circular farmland, and industrial area, is higher than BebyGAN by over 0.3 dB, and ESTUGAN achieves the lowest LPIPS in all scenes, which means the predicted images generated by our method have the optimum visual effect. It also proves that our method can be fine-tuned for different scenes to faithfully reconstruct the actual image distribution.

4.4.2. Qualitative Comparison

We also performed a qualitative comparison to verify the effectiveness of ESTUGAN, as shown in Figure 6. Compared to SRGAN, ESRGAN, and BebyGAN, our proposed method generates more accurate structure information and minimum artifacts, especially in the flat areas. We also reconstruct sharper and more detailed results compared to PSNR-based methods. The effectiveness of our method is well proven.

Table 1. Qualitative comparison of PSNR, SSIM, and LPIPS in the NWPU-RESISC45, the UCMerced dataset, the RSCNN7 dataset, and the DOTA dataset at a four-time scale factor.

Method	NWPU-RESISC45			UCMerced		
	PSNR	**SSIM**	**LPIPS**	**PSNR**	**SSIM**	**LPIPS**
Bicubic	27.61	0.697	0.528	26.96	0.698	0.492
RCAN	29.23	0.772	0.346	28.86	0.776	0.318
RRDB	29.20	0.770	0.362	28.86	0.775	0.338
SwinIR	29.42	0.779	0.340	29.17	0.787	0.312
ESTN (ours)	29.39	0.777	0.341	29.11	0.785	0.312
SRGAN	25.26	0.644	0.233	24.13	0.645	0.258
ESRGAN	26.18	0.711	0.263	25.24	0.717	0.259
BebyGAN	27.80	0.718	0.261	27.28	0.724	0.257
ESTUGAN (ours)	28.12	0.725	0.204	27.81	0.739	0.208
Method	**RSCNN7**			**DOTA**		
	PSNR	**SSIM**	**LPIPS**	**PSNR**	**SSIM**	**LPIPS**
Bicubic	27.99	0.684	0.592	30.89	0.809	0.431
RCAN	29.17	0.744	0.441	33.66	0.868	0.267
RRDB	29.16	0.744	0.449	33.65	0.868	0.273
SwinIR	29.33	0.751	0.436	33.98	0.873	0.264
ESTN (ours)	29.30	0.749	0.438	33.95	0.872	0.266
SRGAN	25.23	0.608	0.284	26.31	0.732	0.272
ESRGAN	26.24	0.692	0.331	28.19	0.824	0.246
BebyGAN	28.07	0.698	0.318	31.06	0.829	0.233
ESTUGAN (ours)	28.15	0.699	0.271	32.17	0.829	0.179

4.5. Ablation Study

We conducted ablation experiments on the test set to verify the performance of the proposed components. In order to verify the performance of the U-Net discriminator, we adopted BebyGAN and ESTUGAN as the baseline to test their performance with the U-Net discriminator and regular discriminator [14,15] respectively. As shown in Figure 7, after adopting the U-Net discriminator, the PSNR of BebyGAN improves by 0.22 dB, the LPIPS decreases by 0.012, and the SSIM increases by 0.001 dB. When replacing the U-Net discriminator with a regular discriminator in the proposed method, the PSNR drops by 0.146 dB, the LPIPS rises 0.008, and the SSIM decreases by nearly 0.01 dB, which significantly affects the reconstruction performance. This shows that the U-Net discriminator provides a more robust ability to identify authenticity. Meanwhile, we visualized the results of the discriminator determination, and the results are shown in Figure 8c, where the black pixels denote that the discriminator makes a negative judgment, while the white pixels indicate that the discriminator generates a positive judgment. Such an accurate pixel-by-pixel judgment facilitates the generator to produce better results for the LPIPS.

In addition, we also verified the effectiveness of the BB loss and the region-aware learning strategy in our approach, as shown in Table 3. Due to the elimination of the BB loss, the performance decreases on both test sets. Similarly, the PSNR, SSIM, and LPIPS deteriorate after the removal the region-aware strategy. It is noteworthy that the performance of the model without the BB loss and the region-aware learning strategy deteriorates more significantly on the UCMerced test set than on the NWPU-RESISC45 dataset. This observation underscores the potential benefits of incorporating the BB loss and the region-aware learning strategy to enhance the model generalizability.

Finally, to demonstrate the performance of our improved deep feature extraction module in the generator, we compared it with two baselines which have the same deep feature extraction module as HAT [12]. We set the number of channels to sixty and the ESTB to four (denoted as baseline1) and six (denoted as baseline2), respectively. Table 4 records the comparison results of our ESTN with two baselines on the UCMerced dataset. As can be seen from the experimental results, neither of the two baselines perform as well as our network. Although baseline2 has a deeper network structure, the effect is not better than baseline1. This proves that the residual structure will suffer performance degradation in the long-term feature extraction phase, and that cascading between residual structures will improve the performance of the remote sensing image SR task.

Table 2. SR results for each class in the NWPU-RESISC45 dataset at a four-time scale factor.

Scene Class	Bicubic PSNR/LPIPS	RCAN PSNR/LPIPS	SwinIR PSNR/LPIPS	SRGAN PSNR/LPIPS	ESRGAN PSNR/LPIPS	BebyGAN PSNR/LPIPS	ESTN (Ours) PSNR/LPIPS	ESTGAN (Ours) PSNR/LPIPS
Airplane	28.57/0.434	31.05/0.229	31.42/0.225	26.84/0.168	26.15/0.207	29.24/0.207	31.32/0.224	29.99/0.153
Airport	27.83/0.551	29.13/0.394	29.26/0.393	25.63/0.244	25.90/0.284	28.01/0.287	29.22/0.395	28.17/0.242
Baseball diamond	27.69/0.520	29.69/0.314	29.88/0.312	26.23/0.201	26.81/0.226	28.44/0.238	29.85/0.312	28.40/0.175
Basketball court	26.53/0.510	28.74/0.276	29.06/0.265	25.75/0.205	26.18/0.253	27.40/0.249	29.01/0.264	27.37/0.176
Beach	30.05/0.485	31.30/0.350	31.36/0.346	27.13/0.205	26.13/0.281	29.23/0.275	31.36/0.347	30.26/0.227
Bridge	29.04/0.450	31.05/0.265	31.23/0.258	28.18/0.169	28.23/0.212	29.56/0.224	31.21/0.259	29.81/0.166
Chaparral	25.54/0.533	27.14/0.324	27.31/0.327	20.74/0.329	24.59/0.233	25.82/0.247	27.31/0.324	25.74/0.158
Church	24.49/0.568	26.39/0.321	26.60/0.315	23.50/0.207	24.43/0.264	25.28/0.276	26.57/0.318	25.25/0.194
Circular farmland	31.21/0.448	33.26/0.246	33.43/0.242	29.99/0.150	28.32/0.197	31.52/0.202	33.41/0.241	32.12/0.153
Cloud	34.81/0.362	36.21/0.267	36.38/0.275	29.07/0.168	27.92/0.159	32.37/0.164	36.35/0.274	34.73/0.153
Commercial area	25.98/0.576	27.53/0.341	27.67/0.334	24.45/0.239	25.38/0.275	26.50/0.282	27.66/0.333	26.51/0.207
Dense residential	22.43/0.660	23.84/0.411	24.01/0.391	20.20/0.257	22.56/0.286	23.02/0.294	24.00/0.391	22.87/0.206
Desert	32.17/0.472	33.13/0.361	33.23/0.361	27.03/0.214	25.05/0.289	30.76/0.270	33.26/0.359	32.08/0.230
Forest	28.47/0.653	29.00/0.553	29.03/0.547	22.33/0.392	27.79/0.358	28.11/0.319	29.04/0.546	27.64/0.288
Freeway	27.34/0.544	28.79/0.350	29.16/0.333	25.71/0.235	26.81/0.266	27.78/0.266	29.06/0.335	27.84/0.198
Golf course	29.26/0.531	31.11/0.340	31.20/0.342	27.53/0.188	28.56/0.243	29.81/0.259	31.20/0.340	29.83/0.188
Ground track field	27.22/0.520	28.89/0.327	29.19/0.318	24.99/0.203	26.55/0.230	27.64/0.239	29.10/0.321	27.75/0.168
Harbor	21.44/0.534	22.91/0.309	23.33/0.273	20.25/0.177	21.83/0.224	22.14/0.228	23.22/0.281	22.04/0.171
Industrial area	27.04/0.509	28.88/0.315	29.09/0.316	24.77/0.198	25.75/0.237	27.35/0.246	29.04/0.315	27.77/0.188
Intersection	23.44/0.587	25.19/0.340	25.38/0.323	22.50/0.269	23.19/0.306	24.02/0.308	25.43/0.327	24.29/0.226
Island	36.18/0.283	37.43/0.189	37.64/0.187	33.43/0.124	29.52/0.158	32.35/0.160	37.67/0.187	35.94/0.124
Lake	30.65/0.495	31.78/0.377	31.84/0.377	26.27/0.262	28.62/0.265	30.19/0.259	31.84/0.378	30.65/0.233
Meadow	29.36/0.675	29.61/0.610	29.63/0.608	24.80/0.357	28.43/0.480	28.92/0.366	29.63/0.605	28.62/0.363
Medium residential	27.45/0.639	28.67/0.442	28.77/0.436	24.97/0.267	26.95/0.310	27.73/0.341	28.74/0.435	27.53/0.242
Mobile home park	22.76/0.660	24.62/0.407	24.84/0.395	21.56/0.236	23.18/0.313	23.69/0.340	24.81/0.393	23.60/0.235
Mountain	29.70/0.555	30.55/0.444	30.60/0.444	26.74/0.268	27.23/0.290	29.45/0.293	30.60/0.443	29.54/0.257
Overpass	27.71/0.515	29.66/0.329	29.83/0.321	27.18/0.199	27.16/0.247	28.48/0.260	29.82/0.325	28.61/0.188
Palace	26.34/0.533	28.11/0.338	28.31/0.333	23.38/0.255	25.75/0.237	26.94/0.237	28.29/0.337	27.03/0.185
Parking lot	21.36/0.579	23.08/0.326	23.46/0.301	20.13/0.229	21.57/0.271	21.93/0.271	23.35/0.301	22.30/0.215
Railway	26.98/0.569	28.37/0.376	28.57/0.366	25.76/0.219	26.51/0.286	27.42/0.289	28.48/0.372	27.35/0.209
Railway station	25.95/0.547	27.65/0.367	27.93/0.360	24.76/0.212	25.13/0.258	26.49/0.253	27.91/0.361	26.82/0.203
Rectangular farmland	31.36/0.542	32.78/0.361	32.93/0.356	30.50/0.214	28.86/0.296	31.16/0.291	32.92/0.358	31.72/0.241
River	29.20/0.482	30.98/0.302	31.13/0.297	27.97/0.176	27.65/0.210	29.53/0.214	31.12/0.298	29.78/0.177
Roundabout	24.97/0.572	26.41/0.385	26.57/0.382	23.75/0.244	24.41/0.283	25.58/0.293	26.56/0.381	25.51/0.226
Runway	29.35/0.437	33.07/0.240	33.87/0.234	29.01/0.171	27.63/0.216	30.71/0.218	33.63/0.235	31.95/0.157
Sea ice	29.60/0.447	31.38/0.303	31.49/0.298	22.89/0.378	27.94/0.221	29.65/0.221	31.47/0.302	30.14/0.182
Ship	27.67/0.494	29.58/0.292	29.79/0.281	26.34/0.203	26.56/0.266	28.25/0.261	29.72/0.286	28.44/0.187
Snowberg	23.89/0.550	25.10/0.408	25.22/0.400	19.42/0.383	23.05/0.272	24.19/0.280	25.22/0.398	24.06/0.225
Sparse residential	26.94/0.657	27.95/0.506	28.08/0.505	24.57/0.319	26.10/0.395	27.15/0.391	28.04/0.503	26.98/0.313
Stadium	26.70/0.506	28.49/0.326	28.65/0.325	24.49/0.223	25.32/0.228	27.15/0.239	28.61/0.324	27.44/0.183
Storage tank	25.72/0.494	27.94/0.282	28.17/0.278	24.98/0.181	25.12/0.204	26.80/0.219	28.12/0.277	26.84/0.154
Tennis court	25.65/0.601	27.51/0.373	27.63/0.366	24.22/0.223	25.31/0.281	26.27/0.295	27.63/0.370	26.33/0.207
Terrace	28.79/0.475	30.49/0.287	30.66/0.283	27.38/0.183	26.73/0.242	29.00/0.256	30.62/0.283	29.42/0.186
Thermal power station	26.60/0.511	28.51/0.315	28.71/0.312	25.13/0.214	25.52/0.222	27.07/0.234	28.66/0.310	27.46/0.186
Wetland	31.14/0.512	32.25/0.370	32.35/0.368	24.51/0.329	29.62/0.268	30.82/0.284	32.36/0.367	30.92/0.226
Mean	27.61/0.528	29.23/0.346	29.42/0.340	25.26/0.233	26.18/0.263	27.80/0.261	29.39/0.341	28.12/0.204
Standard deviation	3.07/0.076	3.03/0.078	3.02/0.079	2.87/0.062	1.94/0.055	2.50/0.046	3.03/0.078	2.94/0.044

Table 3. The comparison of ablation studies on BB loss and region aware strategies in the NWPU-RESISC45 dataset. "Ours" means our proposed ESTUGAN, "w/o BBL" and "w/o RA" indicate the model removing BB loss and the mode removing the region aware strategy.

Dataset	Metrics	Ours	w/o BBL	w/o RA
NWPU-RESISC45	PSNR	28.12	27.95	27.98
	SSIM	0.725	0.717	0.719
	LPIPS	0.204	0.213	0.212
UC-Merced	PSNR	27.81	27.46	27.50
	SSIM	0.739	0.726	0.728
	LPIPS	0.208	0.217	0.215

Table 4. Comparison of using different generator frameworks on the UCMerced dataset.

Generator Settings	PSNR	SSIM	LPIPS
Baseline1	29.04	0.783	0.311
Baseline2	28.60	0.768	0.330
ESTN (ours)	29.11	0.785	0.312

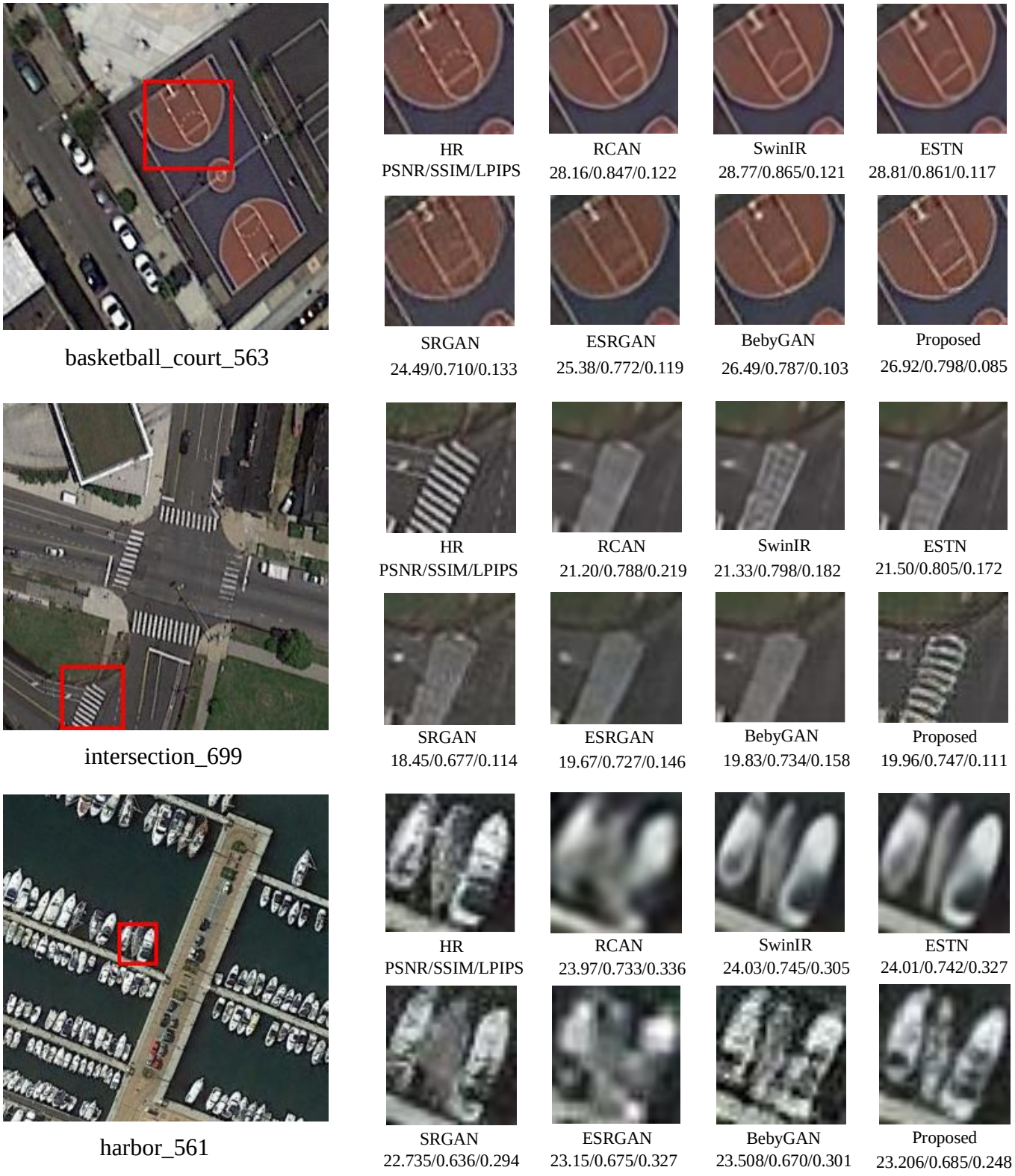


Figure 6. Visualization comparison of various algorithms on the NWPU-RESISC45 dataset with a scale factor $\times 4$.

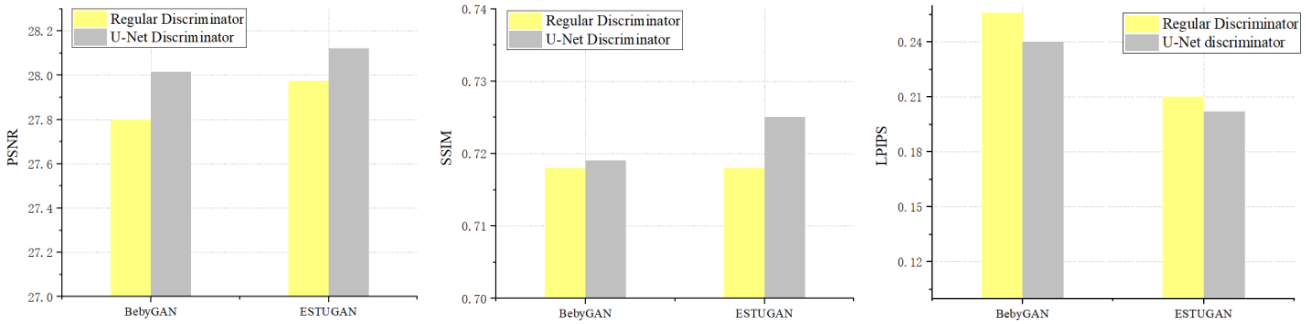
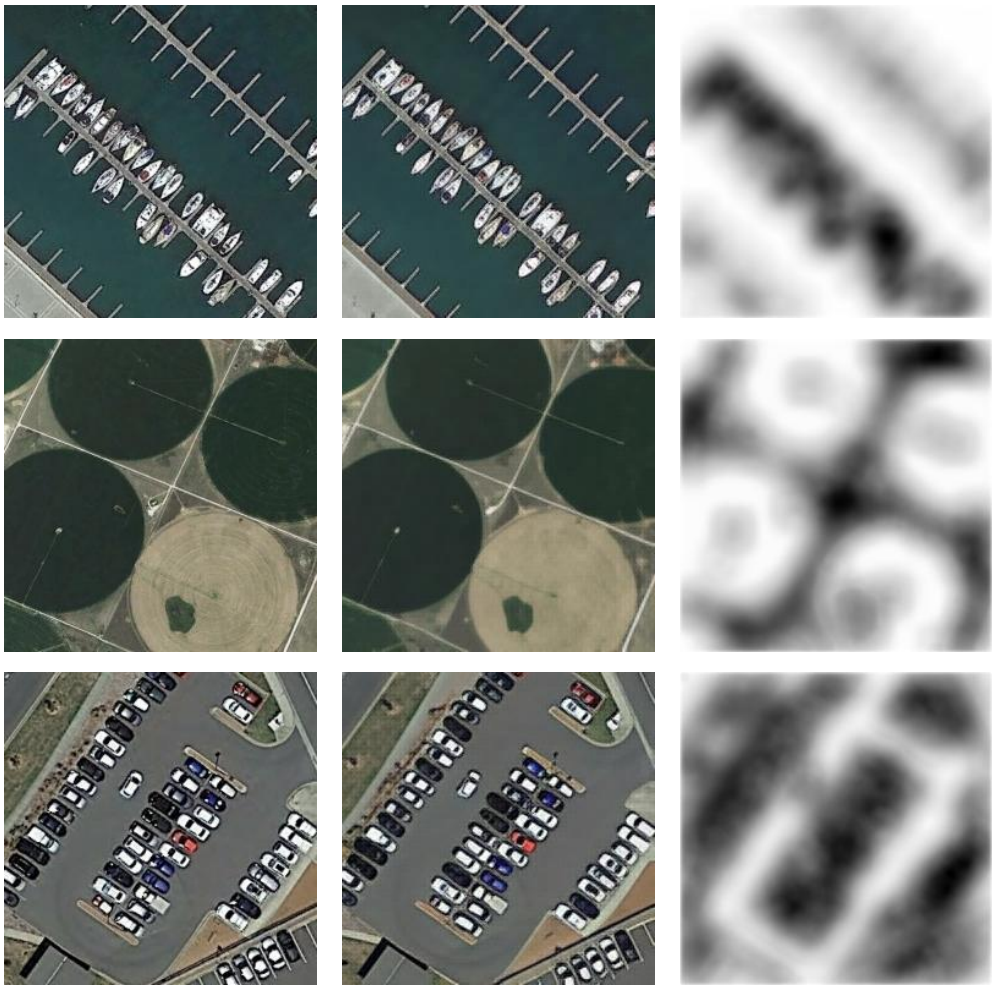


Figure 7. The performance of our proposed ESTUGAN and the BebyGAN on the NWPU-RESISC45 dataset when different discriminators are employed. The discriminators were measured using PSNR, SSIM, and LPIPS metrics.



(a) (b) (c)

Figure 8. The visualization of the U-Net discriminator. (**a**) The original images in the selected dataset. (**b**) The generated images of the proposed generator. (**c**) The discrimination on the generated images.

4.6. Model Complexity Analysis

Figure 9 visualizes the measurement between the parameters and the PSNR of EDSR [64], RCAN [5], RRDB [15], SwinIR [11], HSENet [24], SWCG [65], Resnet [2], and our ESTN. It can be seen that the ESTN is comparable to SwinIR in terms of performance and has an absolute advantage in the parameters, saving over nine parameters (M) compared to SwinIR. Our ESTN performs impressively in terms of the PSNR performance and the number of

parameters. Table 5 comprehensively shows the parameters, FLOPs, and inference time for different methods.

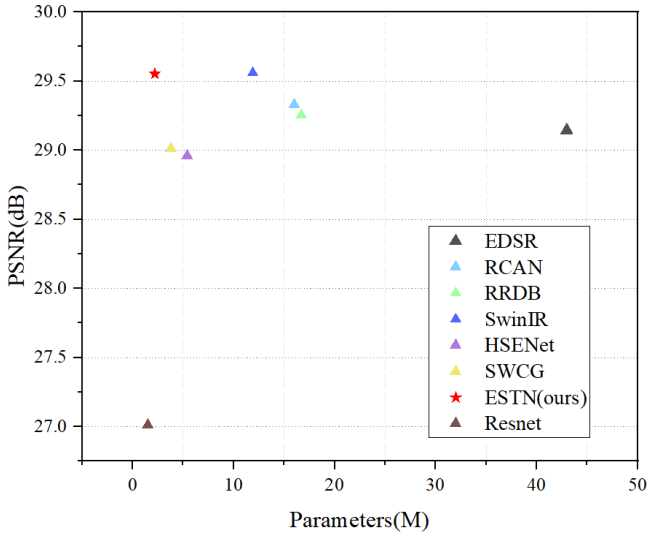


Figure 9. Visualization comparison of parametric quantities and PSNR of diverse approaches.

Table 5. Parameters, FLOPs, and GPU runtime for various super-resolution models. GPU runtime is tested on the Tesla V100 GPU and the input size is 125 × 125.

Model	Parameters	FLOPs	GPU Runtime
RCAN	16 M	233.8 G	0.189 s
RRDB	16.7 M	257.5 G	0.101 s
HSENet	5.4 M	73.3 G	0.155 s
SwinIR	11.9 M	202.2 G	0.288 s
ESTN (ours)	2.2 M	53.5 G	0.165 s

5. Conclusions

In this paper, ESTUGAN was proposed for characteristics of remote sensing images. The generator was the ESTN with the backbone of the Swin Transformer, which combines the advantages of CNN- and transformer-based models, possessing a more powerful expression ability. Meanwhile, the U-Net discriminator with the region-aware learning strategy and the loss strategy that can supervise flexibility was proposed; it effectively suppressed artifacts and guided the generator to recover authentic high-frequency information. Extensive experiments proved that ESTUGAN outperforms existing methods with fewer parameters for remote sensing image SR. Specifically, we tested the performance of our model on four widely used remote sensing datasets. And for the proposed method, sufficient ablation tests were conducted to verify the validity of the components. At the same time, we also explored the network length and the performance of the image SR task to some extent; we found that just adding more functional blocks and increasing the number of parameters does not improve the overall performance, and even decreases it in some specific scenarios.

In the future, we will continue to explore the effectiveness of lightweight models for SR tasks in remote sensing images.

Author Contributions: Conceptualization, L.H.; methodology, L.H.; software, L.H. and T.P.; validation, C.Y. and L.H.; formal analysis, C.Y., L.H. and T.P.; investigation, L.H.; resources, C.Y. and L.H.; data curation, L.H.; writing-original draft preparation, L.H. and T.P.; writing-review and editing, C.Y., L.H., T.P. and Y.L.; visualization, L.H.; supervision, C.Y., Y.L. and T.L.; project administration, L.H.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Liaoning Provincial Science and Technology Department under Grant 2022JH2/101300247 and in part by the Shenyang Municipal Natural Science Foundation under Grant 23-503-6-18.

Data Availability Statement: Not applicable.

Acknowledgments: We are very grateful to the editors and reviewers for their valuable comments, to the providers of all the data used in the paper, and to the people who helped to complete this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dong, C.; Loy, C.C.; He, K.; Tang, X. Learning a Deep Convolutional Network for Image Super-Resolution. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 184–199.
2. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1646–1654.
3. Kim, J.; Lee, J.K.; Lee, K.M. Deeply-Recursive Convolutional Network for Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1637–1645.
4. Lai, W.S.; Huang, J.B.; Ahuja, N.; Yang, M.H. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 26 June–21 July 2017; pp. 624–632.
5. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image Super-Resolution Using Very Deep Residual Channel Attention Networks. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
6. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30, Long Beach, CA, USA, 4–9 December 2017.
7. Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; Shlens, J. Scaling Local Self-Attention for Parameter Efficient Visual Backbones. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Kuala Lumpur, Malaysia, 18–20 December 2021; pp. 1637–1645.
8. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H.U. A general u-shaped transformer for image restoration. *arXiv* **2021**, arXiv:2106.03106.
9. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
10. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
11. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Gool, L.V.; Timofte, R. Swinir: Image Restoration Using Swin Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1833–1844.
12. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating More Pixels in Image Super-Resolution Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 22367–22377.
13. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2014**, *63*, 139–144. [CrossRef]
14. Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
15. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. Esrgan: Enhanced Super-Resolution Generative Adversarial Networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
16. Zhang, W.; Liu, Y.; Dong, C.; Qiao, Y. Ranksrgan: Generative Adversarial Networks with Ranker for Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3096–3105.
17. Li, W.; Zhou, K.; Qi, L.; Lu, L.; Lu, J. Best-Buddy Gans for Highly Detailed Image Super-Resolution. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 2022; pp. 1412–1420.
18. Lei, S.; Shi, Z.; Zou, Z. Super-resolution for remote sensing images via local–global combined network. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1243–1247. [CrossRef]
19. Jiang, K.; Wang, Z.; Yi, P.; Wang, G.; Lu, T.; Jiang, J. Edge-enhanced GAN for remote sensing image superresolution. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5799–5812. [CrossRef]
20. Pan, Z.; Ma, W.; Guo, J.; Lei, B. Super-resolution of single remote sensing image based on residual dense backprojection networks. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7918–7933. [CrossRef]
21. Ma, W.; Pan, Z.; Guo, J.; Lei, B. Achieving super-resolution remote sensing images via the wavelet transform combined with the recursive resnet. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 3512–3527. [CrossRef]

22. Jiang, W.; Zhao, L.; Wang, Y.J.; Liu, W.; Liu, B.D. U-shaped attention connection network for remote-sensing image super-resolution. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
23. Wu, H.; Zhang, L.; Ma, J. Remote sensing image super-resolution via saliency-guided feedback GANs. *IEEE Trans. Geosci. Remote Sens.* **2020**, *60*, 1–16. [CrossRef]
24. Lei, S.; Shi, Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–10. [CrossRef]
25. Liu, Z.; Feng, R.; Wang, L.; Han, W.; Zeng, T. Dual learning-based graph neural network for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–14. [CrossRef]
26. Yu, Y.; Li, X.; Liu, F. E-DBPN: Enhanced deep back-projection networks for remote sensing scene image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5503–5515. [CrossRef]
27. Jia, S.; Wang, Z.; Li, Q.; Jia, X.; Xu, M. Multiattention generative adversarial network for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
28. Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Guo, B. Swin Transformer v2: Scaling up Capacity and Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 12009–12019.
29. Choi, H.; Lee, J.; Yang, J. N-Gram in Swin Transformers for Efficient Lightweight Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Vancouver, BC, Canada, 18–22 June 2023; pp. 2071–2081.
30. Hassani, A.; Shi, H. Dilated neighborhood attention transformer. *arXiv* **2022**, arXiv:2209.15001.
31. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and Improving the Image Quality of Stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–20 June 2020; pp. 8110–8119.
32. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. *arXiv* **2018**, arXiv:1809.11096.
33. Karras, T.; Laine, S.; Aila, T. A Style-Based Generator Architecture for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4401–4410.
34. Yang, M.; Sowmya, A. New image quality evaluation metric for underwater video. *IEEE Signal Process. Lett.* **2014**, *21*, 1215–1219. [CrossRef]
35. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
36. Fuoli, D.; Van Gool, L.; Timofte, R. Fourier Space Losses for Efficient Perceptual Image Super-Resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2360–2369.
37. Liang, J.; Zeng, H.; Zhang, L. Details or Artifacts: A Locally Discriminative Learning Approach to Realistic Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5657–5666.
38. Jiang, K.; Wang, Z.; Yi, P.; Jiang, J.; Xiao, J.; Yao, Y. Deep distillation recursive network for remote sensing imagery super-resolution. *Remote Sens.* **2018**, *10*, 1700. [CrossRef]
39. Zhang, D.; Shao, J.; Li, X.; Shen, H.T. Remote sensing image super-resolution via mixed high-order attention network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 5183–5196. [CrossRef]
40. Zhang, S.; Yuan, Q.; Li, J.; Sun, J.; Zhang, X. Scene-adaptive remote sensing image super-resolution using a multiscale attention network. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 4764–4779. [CrossRef]
41. Li, Y.; Mavromatis, S.; Zhang, F.; Du, Z.; Sequeira, J.; Wang, Z.; Zhao, X.; Liu, R. Single-image super-resolution for remote sensing images using a deep generative adversarial network with local and global attention mechanisms. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–24. [CrossRef]
42. Huynh-Thu, Q.; Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electron. Lett.* **2008**, *44*, 800–801. [CrossRef]
43. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
44. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef]
45. Zhou, W.; Wang, Z.; Chen, Z. Image Super-Resolution Quality Assessment: Structural Fidelity versus Statistical Naturalness. In Proceedings of the IEEE International Conference on Quality of Multimedia Experience (QoMEX), Virtual, 14–17 June 2021; pp. 61–64.
46. Zhou, W.; Wang, Z. Quality Assessment of Image Super-Resolution: Balancing Deterministic and Statistical Fidelity. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 934–942.
47. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 586–595.
48. Zhou, W.; Jiang, Q.; Wang, Y.; Chen, Z.; Li, W. Blind quality assessment for image superresolution using deep two-stream convolutional networks. *Inf. Sci.* **2020**, *528*, 205–218. [CrossRef]

49. Chen, T.; Liu, H.; Ma, Z.; Shen, Q.; Cao, X.; Wang, Y. End-to-end learnt image compression via non-local attention optimization and improved context modeling. *IEEE Trans. Image Process.* **2021**, *30*, 3179–3191. [CrossRef]
50. Zhang, Y.; Li, K.; Li, K.; Zhong, B.; Fu, Y. Residual non-local attention networks for image restoration. *arXiv* **2019**, arXiv:1903.10082.
51. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883.
52. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; pp. 234–241.
53. Schonfeld, E.; Schiele, B.; Khoreva, A. A U-Net Based Discriminator for Generative Adversarial Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8207–8216.
54. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv* **2018**, arXiv:1802.05957.
55. Kindermann, S.; Osher, S.; Jones, P.W. Deblurring and denoising of images by nonlocal functionals. *Multiscale Model. Simul.* **2005**, *4*, 1091–1115. [CrossRef]
56. Protter, M.; Elad, M.; Takeda, H.; Milanfar, P. Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. Image Process.* **2008**, *18*, 36–51. [CrossRef]
57. Pan, T.; Zhang, L.; Song, Y.; Liu, Y. Hybrid Attention Compression Network with Light Graph Attention Module for Remote Sensing images. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
58. Glasner, D.; Bagon, S.; Irani, M. Super-resolution from a single image. In Proceedings of the IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 349–356.
59. Huang, J.B.; Singh, A.; Ahuja, N. Single Image Super-Resolution from Transformed Self-Exemplars. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 8–12 June 2015; pp. 5197–5206.
60. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
61. Yang, Y.; Newsam, S. Bag-of-Visual-Words and Spatial Extensions for Land-Use Classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
62. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning-based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2015**, *12*, 2321–2325. [CrossRef]
63. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
64. Lim, B.; Son, S.; Kim, H.; Nah, S.; Mu Lee, K. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 26 June–21 July 2017; pp. 136–144.
65. Tu, J.; Mei, G.; Ma, Z.; Piccialli, F. SWCGAN: Generative adversarial network combining swin transformer and CNN for remote sensing image super-resolution. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 5662–5673. [CrossRef]