remote sensing

MDPI

Article

SA-GAN: A Second Order Attention Generator Adversarial Network with Region Aware Strategy for Real Satellite Images Super Resolution Reconstruction

Jiayi Zhao [1,2], Yong Ma [1], Fu Chen [1,*], Erping Shang [1], Wutao Yao [1], Shuyan Zhang [1] and Jin Yang [1]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
* Correspondence: chenfu@aircas.ac.cn

Abstract: High-resolution (HR) remote sensing images have important applications in many scenarios, and improving the resolution of remote sensing images via algorithms is one of the key research fields. However, current super-resolution (SR) algorithms, which are trained on synthetic datasets, tend to have poor performance in real-world low-resolution (LR) images. Moreover, due to the inherent complexity of real-world remote sensing images, current models are prone to color distortion, blurred edges, and unrealistic artifacts. To address these issues, real-SR datasets using the Gao Fen (GF) satellite images at different spatial resolutions have been established to simulate real degradation situations; moreover, a second-order attention generator adversarial attention network (SA-GAN) model based on real-world remote sensing images is proposed to implement the SR task. In the generator network, a second-order channel attention mechanism and a region-level non-local module are used to fully utilize the a priori information in low-resolution (LR) images, as well as adopting region-aware loss to suppress artifact generation. Experiments on test data demonstrate that the model delivers good performance for quantitative metrics, and the visual quality outperforms that of previous approaches. The Frechet inception distance score (FID) and the learned perceptual image patch similarity (LPIPS) value using the proposed method are improved by 17.67% and 6.61%, respectively. Migration experiments in real scenarios also demonstrate the effectiveness and robustness of the method.

Keywords: super-resolution; region aware; second-order channel attention; Gao Fen satellite; region-level non-local

Citation: Zhao, J.; Ma, Y.; Chen, F.; Shang, E.; Yao, W.; Zhang, S.; Yang, J. SA-GAN: A Second Order Attention Generator Adversarial Network with Region Aware Strategy for Real Satellite Images Super Resolution Reconstruction. *Remote Sens.* **2023**, *15*, 1391. https://doi.org/10.3390/rs15051391

Academic Editors: Igor Yanovsky and Jing Qin

Received: 28 December 2022
Revised: 24 February 2023
Accepted: 24 February 2023
Published: 1 March 2023

1. Introduction

High-resolution (HR) remote sensing images provide detailed texture information of ground objects, which are essential for many applications, such as the classification of land cover [1], object detection [2], building extraction [3], and change detection [4]. However, the spatial resolution of remote sensing images is influenced by the sensor hardware and environmental factors [5], and it is relatively difficult to obtain HR images at a specific time. At the hardware level, it is possible to increase the number of satellites to provide more HR satellites or directly improve the production technology of sensors to directly improve the resolution. These options tend to be more costly in most instances. In comparison with the above strategies, super-resolution (SR) image technology is more convenient and of relatively low cost. SR is a technique for generating HR images from low-resolution (LR) images. The approach can be categorized into the single-image super-resolution (SISR) or multi-image super-resolution (MISR). Although the multi-image technique can provide more a priori information, it is difficult to obtain multiple remote sensing images of the same scene.

The traditional image SR algorithms can be grouped into two main categories, interpolation-based algorithms, and reconstruction-based algorithms. The interpolation-based image

SR methods reconstruct HR images by computing the pixels of the point to be sought from the known pixel values around the point to be interpolated. The main commonly used interpolation algorithms include the nearest neighbor interpolation [6], the bilinear interpolation [7], and the bicubic interpolation [8] methods. Interpolation algorithms tend to be faster and simpler than other methods. However, linear model algorithms have limited ability to recover high-frequency detail. Reconstruction-based algorithms use complex a priori knowledge as constraints to reconstruct the HR images, such as iterative back projection (IBP) [9], projection onto convex sets (POCS) [10], and maximum-a-posteriori (MAP) approach [11]. Although reconstruction-based algorithms utilize a priori information, they do not always generate acceptable results for complex images.

In recent years, with the development of deep learning techniques, a series of deep learning-based methods have emerged in the field of SR. Dong et al. [12] proposed a super-resolution convolutional neural network (SRCNN), which learns the mapping relationship between bicubic linear interpolation images and HR images through the use of neural networks. Although the SRCNN can outperform traditional-based methods, bicubic LR images are computationally slow in the network. To alleviate the problem, a fast super-resolution convolutional neural network (FSRCNN) [13] has added transposed convolution operations to the network to reduce the computational time of the network. These networks are shallow, and their performance is affected by the depth of the network; however, increasing the depth of the network can lead to gradient explosion and gradient disappearance. To deepen the depth of the network and obtain a stronger learning ability, very deep super-resolution (VDSR) [14] incorporates residual learning and gradient cropping to mitigate the network gradient explosion-disappearance problem. Further, the use of the deeply recursive convolutional network (DRCN) [15] increases the network depth skip connections and recursive supervision. Additionally, the deep recursive residual network (DRRN) [16] improves the performance of the network by proposing recursive learning of local residual connections and residual units on the basis of the DRCN. Moreover, enhanced deep super-resolution (EDSR) [17] and SRResNet [18], which use residual connections [19], deepen the network depth and avoid the gradient problem. EDSR removes the batch normalization (BN) [20] layer and uses a residual scaling module to increase the stability of the training.

In previously reported networks, all the different channels are characterized and treated equally, and the residual channel attention network (RCAN) [21] uses channel attention to enhance the ability of the network to distinguish between the different channels. The dual regression network (DRN) [22] proposes a dual regression scheme by introducing an additional constraint on the LR data to reduce the space of the possible functions. The local texture estimator (LTE) [23] proposes an LTE, a dominant-frequency estimator for natural images, enabling an implicit function to capture fine detail while reconstructing images in a continuous manner. To reduce the amount of computation in the SR network, sparse mask SR (SMSR) [24] is adopted to learn sparse masks to prune redundant computation. In addition to the CNN-based models, the transformer [25] model has been used in the field of SR due to its excellent global attention mechanism and texture transformer network (TTSR) [26]. The TTSR uses the LR and the reference image (Ref) as a query and a keyword. Joint feature learning is then performed between the LR and the Ref to extract the relationship between the deep features by global attention, and thus the texture features are displayed accurately. Although the transformer can obtain the global receptive field, the computational effort increases rapidly with image size [27]. Liang et al. [28] used a Swin transformer [29] for deep feature extraction to reduce the computational effort in computing attention. To activate more input pixels for reconstruction, the hybrid attention transformer (HAT) [30] was adopted, combining channel attention and self-attention schemes, which exploit the respective complementary advantages.

CNN and transformer-based methods achieve better recovery results compared to the traditional SR methods, but the recovered images are sometimes too smooth and lack high-frequency detail. Compared to these methods, the generating adversarial network

(GAN) based SR method is able to produce more detailed textures. The GAN consists of a generator and a discriminator. The generator generates an image from the input, the discriminator determines whether the generated image is true or false, and then alternately is optimized to reach the Nash equilibrium [31]. The super-resolution generative adversarial network (SRGAN) [18] uses GAN to solve the SR problem and proposes a perceptual loss to produce more realistic textures. The enhanced super-resolution generative adversarial network (ESRGAN) [32] improves the discriminator using the relativistic average GAN (RaGAN) [33] and introduces the residual-in-residual dense block (RDDB) to improve the model. The spatial feature transforms generative adversarial network (SFTGAN) [34] uses an SFT module to effectively combine the images into the network to improve detailed texture in the GAN networks. The Real-ESRGAN [35], a high-order degradation modeling process, is introduced to better simulate complex real-world degradations and employs a U-Net discriminator with spectral normalization to increase the discriminator capability and stabilize the training dynamics.

In addition to improvements in the model itself, research on the degradation of the HR-LR has been a hot topic in recent years. The super-resolution network for multiple degradations (SRMD) [36] proposes a general framework featuring a dimensionality stretching strategy that enables a single convolutional SR network to take two key factors of the SISR degradation process, that is, the blur kernel and the noise level, as the inputs. A unified dynamic convolutional network (UDVD) [37] introduces a dynamic convolution based on the SRMD and applies dynamic convolution to the up-sampling process. Iterative kernel correction (IKC) [38] solves the artifacts caused by kernel mismatch by correcting the estimated blur kernels through an iterative correction mechanism. Inspired by contrast learning, domain adaptation super resolution (DASR) [39] is proposed as an unsupervised degradation-aware network, which is based on representation learning, to handle different degradation situations adaptively. The unpaired SR [40] proposes a probabilistic degeneracy model (PDM) that studies the degeneracy D as a random variable and learns its distribution by modeling the mapping from a priori random variables Z to D. Blind image super-resolution with elaborate degradation modeling (BSRDM) [41] proposes a patch-based noise model to increase the degrees of freedom of the model for noise representation and to facilitate novel construction of a concise yet effective kernel generator.

The SR has become a research hotspot in image processing in remote sensing due to the huge demand for high spatial resolution in many remote sensing tasks. Jiang et al. [42] proposed the distillation recursive network (DDRN) for video satellite image SR. Galar et al. [43] used the EDSR with several modifications on Sentinel-2 and Planet images. Romero et al. [44] implemented and trained a model based on ESRGAN with pairs of WorldView-Sentinel images to generate a super-resolution multi-spectral Sentinel-2 output with a scaling factor of 5. Zabalza et al. [45] exploited the residual network (SARNet) to increase the spectral resolution of the Sentinel-2 images from the original 10 m to 2.5 m. Karwowska et al. [46] improved the resolution of satellite images acquired with the World View 2 satellite using the ESRGAN network with window functions.

Although the above-mentioned methods deliver good performance, there is still scope for improvement in the remote sensing image SR mission. First, the real degradation of remote sensing images is very complex, given that the process involves the diffraction limit of the lens, disturbances by the atmosphere, the relative movement of the imaging platforms, and the impacts of the different types of imaging noise. All these factors lead to difficulties in producing valid results for real remote sensing images, even for models that consider multiple degradations [47,48]. Therefore, ground truth data are very important for SR in remote sensing images. However, there are few studies on real ground truth data with high spatial resolution, especially at the meter scale. Second, the GAN network can produce more high-frequency information. Nonetheless, adversarial training is unstable and often produces unpleasant visual artifacts, which is a problem exacerbated by the complexity of the remote sensing image distribution. Third, the remote sensing images are complex and contain detailed and rich information. This both generates difficulties in the

SR of remote sensing images and provides more a priori information. The existing GAN networks do not focus on the similarities and differences in this information.

To solve these problems, an SR dataset of real remote sensing images was built using different spatial resolutions for the Gao Fen (GF) satellite data, whereby a GAN model based on second-order channel attention and a region-level non-local module is proposed to utilize the rich a priori information of remote sensing images, and finally, use a region-aware strategy to inhibit the generation of artifacts.

The main contributions of this study are as follows:

1. A real SR dataset based on GF6/1 and GF2/7 satellite data is produced to simulate the degradation process of real remote sensing images.
2. A region-aware strategy is added to the training process to reduce the artifacts generated in GAN and improve the visual quality of the results.
3. An adversarial generative network for SR reconstruction of real remote sensing images was designed, whereby a second-order channel attention mechanism was used to treat channel features differently. A region-level non-local module was added to the generative network to capture long-range dependencies between features, which achieved an accurate restoration of feature structure information.

The paper is organized as follows: The proposed SR method is introduced in detail in Section 2. The evaluation experiments for the different methods are described in Section 3. A further discussion of the proposed method is given in Section 4. Finally, future research directions are specified in the conclusions in Section 5.

2. Materials and Methods

In this section, we describe the overall architecture and specific details of the method, including the region-aware strategy, the network design, and the loss functions. The overall framework of our approach is outlined in Figure 1.

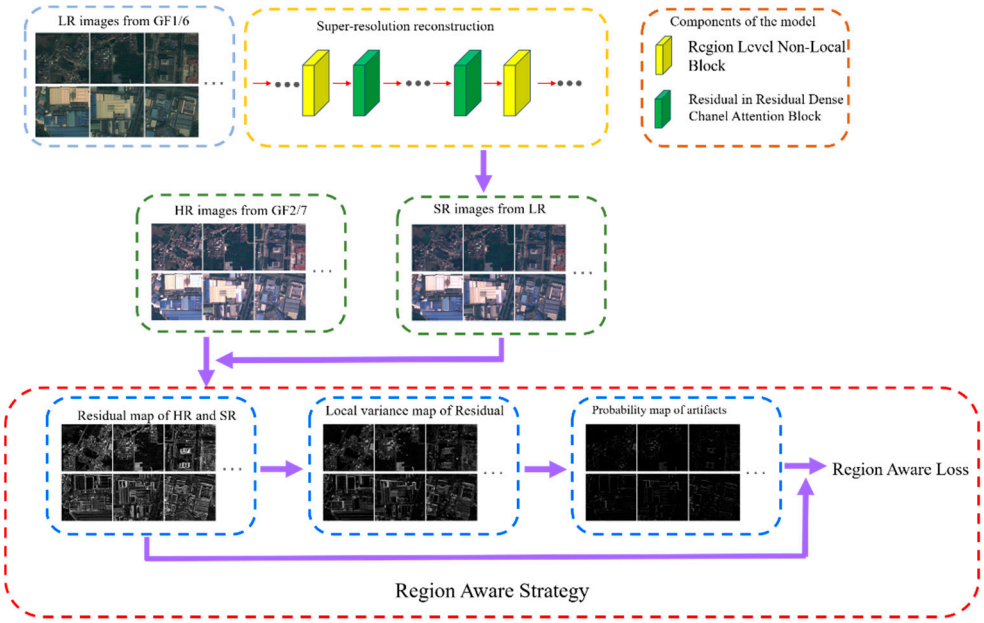


Figure 1. Framework of the method. For the LR input, the reconstruction is performed by a GAN network containing second-order channel attention and region-level non-local blocks; region-aware loss is used to constrain the generation of artifacts.

2.1. Region-Aware Strategy

The GAN generates richer and more detailed information by employing adversarial loss, but learning the mapping from LR to HR images is typically an ill-defined problem given that there exist many HR images that can be downscaled to obtain the same LR

image [49]; this makes GAN generate in perpetuity undesirable artifacts in addition to the details. For smoothed areas, the LR images retain much-structured information of the HR images, and the network can reproduce them with better visual quality; however, for regions rich in high-frequency detail, the large inter-pixel variation makes the SR results produce irregular artifacts. As shown in Figure 2, the farmland in the first row and the dense housing in the second row represent the smooth region and the complex region, respectively. The use of spatial information can reduce the generation of artifacts in GAN networks [50], which is also applicable in remote sensing images. In order to reduce the artifacts generated during the training process, a region-aware strategy is adopted in this study.

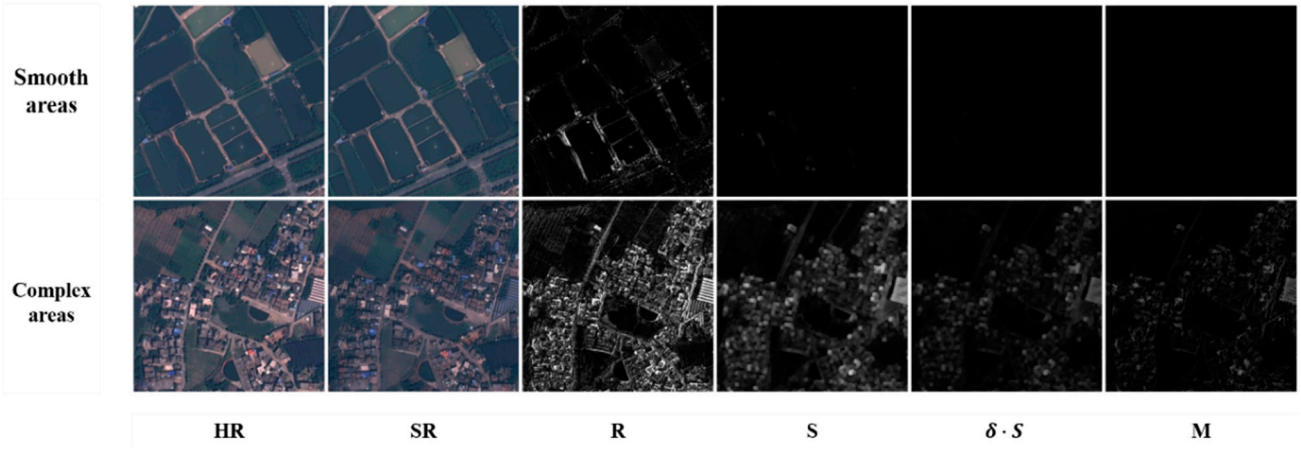


Figure 2. Visualization of the process of generating artifact probability maps. HR, SR, R, S, $\delta \cdot S$ and M represent the outputs of the GAN model, the HR image, the residual map for the HR and SR calculations, the local variance map, the corrected local variance map, and the probability map for the artifacts, respectively.

For a remote-sensing image I_{SR}, the goal is to produce a pixel-wise map $M \in \mathrm{R}^{H \times W \times 1}$, where $M(i,j) \in [0,1]$ indicates the probability of I_{SR} being an artifact pixel. Considering that both artifacts and details are high-frequency image components and there is a better recovery of the network with the smoothed region, the residual between the HR image I_{HR} and the SISR result I_{SR} is first calculated to extract the high-frequency component as follows:

$$R = I_{HR} - I_{SR} \tag{1}$$

As shown in the 3rd column of Figure 2, the residuals are relatively small in the smooth region, while the residuals are large in the high-frequency detail-rich region. However, the region with larger residuals does not necessarily represent artifacts, given that artifacts exist that are not well represented by residuals alone. Under normal circumstances, for regions with large residuals, if there is no drastic change in the residuals, it means that there is only an overall shift in pixel values, which is expressed in the image as a difference in color rather than in structural information; thus it is assumed that the probability is greater that the region is not an artifact. Therefore, the probability of artifacts can be further expressed by calculating the local variance S of the residuals R:

$$S(i,j) = \frac{1}{(n+1)^2} \sum_{x=i-\frac{n}{2}}^{i+\frac{n}{2}} \sum_{y=i-\frac{n}{2}}^{i+\frac{n}{2}} (R(x,y) - \mu) \tag{2}$$

Among them,

$$\mu = \frac{1}{(n+1)} \sum_{x=i-\frac{n}{2}}^{i+\frac{n}{2}} \sum_{y=i-\frac{n}{2}}^{i+\frac{n}{2}} R(x,y) \tag{3}$$

where n represents the size of the local region. We can test the effect of different n values on the local variance, where the artifacts are mostly concentrated in the interior of the building in the SR images. Equation (2) calculates the variance of the $n \times n$ domain at each location in the R. As n increases, the range of the calculated area becomes larger. As shown in Figure 3, when n is small, the area with a larger response cannot completely cover the artifacts. When n is larger, the area with a larger response is larger than the actual area of the artifacts. It can be seen from the figure that when $n = 6$, the artifacts can be better represented in most regions. Therefore, in this study, n is set to 6.

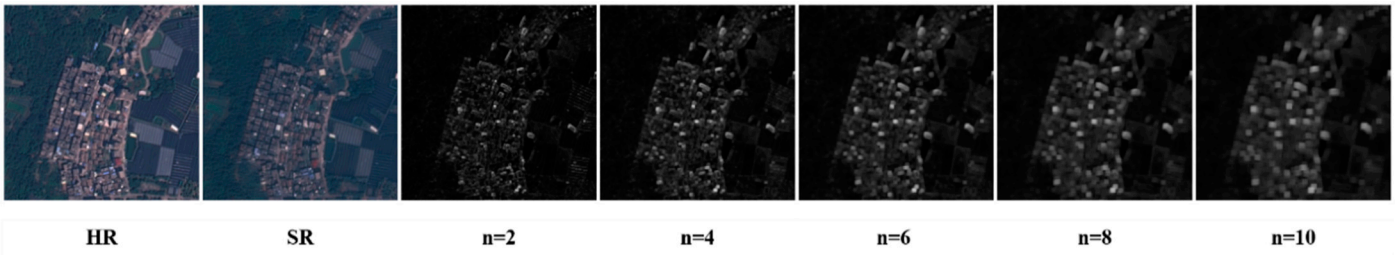


Figure 3. Visualization of the local variance maps at different n values in Equation (2). HR and SR represent the outputs of the GAN model and the HR image, respectively.

As shown in the 4th column of Figure 2, S can indicate artifacts better, but S only considers local information and does not incorporate global information. Therefore, the global variance is calculated as follows:

$$\delta = (var(R))^{\frac{1}{\alpha}} \tag{4}$$

in order to adjust the local variance according to the global variance, where α is a weighting factor and α is larger, the greater the influence of the global method. We fixed α to be $1/4$ through the experiments.

As shown in the 5th column of Figure 2, it can be seen that the probability of artifacts occurring is already almost zero for smooth areas such as farmland. Finally, to address the instability of the GAN network training, the exponential moving average (EMA) approach is used to generate a stable model W^k_{EMA} as follows:

$$W^k_{EMA} = \alpha \cdot W^{k-1}_{EMA} + (1 - \alpha) \cdot W^k \tag{5}$$

where W^k is the K^{th} model, and W^k_{EMA} is the exponential moving average of the model based on calculating the previous K rounds. As in a priori research on EMA [51], we set $\alpha = 0.999$. Compared to W^k, W^k_{EMA} is an integral of multiple models and, therefore, is more stable and generates fewer artifacts, thus W^k_{EMA} is used to highlight the correct direction of the gradient descent of the weight for the W^k through SR results. That is, $I_{SREMA} = W^k_{EMA}(I_{LR})$ and $I_{SR} = W^k(I_{LR})$, and the residual map $R_1 = I_{HR} - I_{SR}$ and $R_2 = I_{HR} - I_{SREMA}$ of the two model results are calculated. If the model can be updated in the correct direction, the generated R_1 of the current model should always be less than R_2, however, the instability of the model leads to the possibility of the model being updated in the wrong direction at locations where the residuals of R_1 are larger than R_2, and, therefore, should not be penalized. Thus, we take the part of R_1 that is larger than R_2 as the artifact to obtain the refined artifact map M:

$$M(i,j) = \begin{cases} 0, & if\ |R_1(i,j)| < |R_2(i,j)|; \\ \delta \cdot S(i,j), & if\ |R_1(i,j)| \geq |R_2(i,j)|. \end{cases} \tag{6}$$

As shown in the 6th column of Figure 3, the improved M retains the realistic details while more accurately representing the artifacts.

2.2. Generative Network

The generative network architecture is shown in Figure 4. The generator network part consists of four main parts: shallow feature extraction, the self-attentive module based on the non-local module, the residual in the residual dense channel attention block (RRDCB), which is based on the deep feature extraction and the reconstruction part. The network details are introduced in the next section.

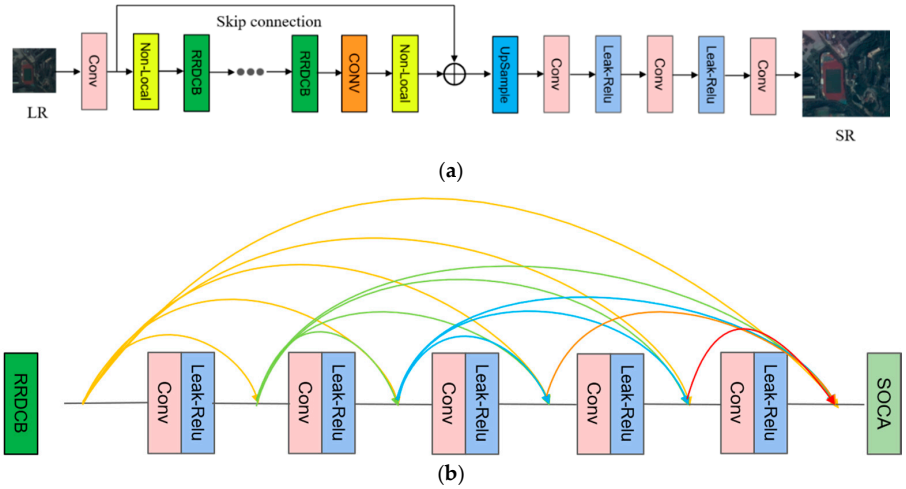


Figure 4. Network structure of generative network: (**a**) The structure of the generator. LR represents low-resolution images, and SR represents the results for network reconstruction. (**b**) The structure of the residual in residual dense second-order channel attention block (RRDCB).

For an LR input $I_{LR} \in R^{H \times W \times C}$, a convolutional layer is used for shallow feature extraction:

$$F_1 = CONV^{k3n64s1}(I_{LR}) \tag{7}$$

where $CONV^{k3n64s1}$ represents a convolutional layer with a convolutional kernel size of 3, a featured channel of 64, and a step size of 1. For the initially proposed features, the region-level non-local module is used to capture the long-distance dependencies to obtain F_{nl} the process is expressed as follows:

$$F_{nl1} = Nonlocal(F_1) \tag{8}$$

Nonlocal represents a self-attentive mechanism at the regional level; more details on *Nonlocal* will be given in Section 2.5.

After the non-local feature F_{nl1} passes through multiple RRDCB blocks, the internal structure of the RRDCB is shown in Figure 4. Each RRDCB includes five convolutional layers and the *Leaky-Relu* activation function. The output of each convolutional layer is concatenated together with the previously passed convolutional layers, and finally, the output passes through the SOCA module and adds the input of the RRDCB, which is multiplied by a weight factor β between 0 and 1 to speed up the training of the model [52]. The process can be expressed as follows:

$$F_{21} = LeakyRelu\left(CONV_{f1}^{k3n32s1}(F_1)\right) \tag{9}$$

$$F_{22} = LeakyRelu\left(CONV_{f2}^{k3n32s1}(Concat(F_1, F_{21}))\right) \tag{10}$$

$$\cdots\cdots$$

$$F_{25} = LeakyRelu\left(CONV_{f5}^{k3n64s1}(Concat(F_1, F_{21}, \ldots F_{24}))\right) \tag{11}$$

$$F_{2soca} = SOCA(F_{25}) * F_{25} \tag{12}$$

$$F_2 = F_1 * \beta + F_{2soca} \tag{13}$$

where *Concat* represents the concatenate operation, $CONV_{fn}^{k3n32s1}$ represents the n^{th} convolutional layer used for feature extraction of size 3×3, the output channel is 32, and the step size is 1, *LeakyRelu* represents the activation function used, SOCA represents the second-order channel attention mechanism, which is described in detail in Section 2.4.

The RRDCB enables the features to be fused at the multi-scale level and to perform depth feature extraction, and captures the long-range dependencies between the depth features again by the region-level non-local operation:

$$F_n = RRDCB_{n-1}(RRDCB_{n-2}(\ldots)) \tag{14}$$

$$F_{nl2} = NonLocal(F_n \tag{15}$$

Finally, the fusion of the shallow features and deep features through a skip connection and bicubic linear interpolation is performed for the improvement of the resolution; afterward, refinement of the features is achieved by a combination of two-layer convolution and *Leak-Relu* and the final SR result is obtained by a convolution layer and fine-tuning:

$$F_{\text{coarse}} = F_1 + F_{nl2} \tag{16}$$

$$F_{refine1} = LeakyRelu(CONV^{k3n64s1}(Upsample(F_{\text{coarse}}))) \tag{17}$$

$$F_{refine2} = LeakyRelu(CONV^{k3n64s1}(F_{refine1})) \tag{18}$$

$$I_{SR} = CONV^{k3n64s1}(F_{refine2}) \tag{19}$$

The $Upsample$ represents the bicubic linear interpolation upsampling, and $I_{SR} \in R^{sH \times sW \times C}$ is the final SR image.

2.3. Discriminator Network

The result of the discriminator network is shown in Figure 5. The role of the discriminator is to determine whether the input is an SR or HR image based on the differences in the distribution of the input image. The input, be it an SR image or an HR image, then passes through a basic block consisting of convolutional layers, Leaky-Relu, and BN layers, in which the convolutional layers are divided into two different step sizes. A convolutional layer with a step size of 1 and kernel size of 3×3 is used for feature extraction, which is a smaller computation than a 5×5 kernel size and a larger perceptual field than a 1×1 size. A convolutional layer with a step size of 2 and kernel size of 4×4 is used to expand the perceptual field of the network and reduce the resolution of the feature map to 1/2 of the original. The whole process of the discriminator may be represented as follows:

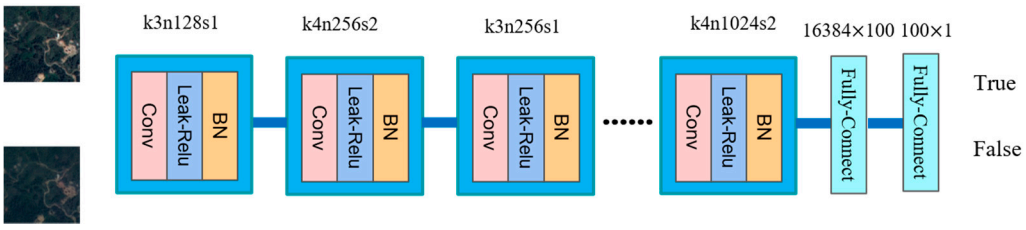


Figure 5. Network Structure of the Discriminator Network.

2.4. Second-Order Channel Attention Mechanism

Prior to Se-Net [53], the CNN mostly did not consider the interdependence between the feature channels, and the Se-Net obtained the first-order statistics of each channel by global pooling and adaptively learned the weights of each feature map by a fully connected layer. Although a differentiated representation of the channels is obtained, higher-order statistics are ignored, affecting the network's differentiation of the importance of different channels.

Some related work [54,55] has shown that second-order statistics in CNN networks are of assistance in learning more discriminative representations. Therefore, in this study, a second-order channel attention module (SOCA) was added to the generative network part to enhance the differentiation ability of the network. The process whereby the second-order channel attention is used in this paper is illustrated in Figure 6. For a feature map of H × W × C $F = [F_1, F_2, \ldots\ldots, F_n]$, F is first reshaped into a matrix X of HW × C, and the covariance matrix COV of X is calculated:

$$X' = X - \overline{X} \tag{20}$$

$$COV = \frac{1}{c} X' X'^T \tag{21}$$

where $\overline{X}$ is the mean value of different channels at the same position. The covariance matrix COV can represent the similarity between channel features. A positive covariance indicates that the two channels have a positive relationship, whereas a negative covariance shows that they have a negative relationship. If two channels do not vary together, then they will display a zero covariance. The properties of the covariance matrix help to distinguish the importance of channel features. Furthermore, normalizing the covariance matrix gives a more differentiated representation of the features [55] due to the semi-positive nature of the COV. The following eigenvalue decomposition can be performed:

$$COV = U \Lambda U^T \tag{22}$$

where Λ is a diagonal matrix of non-decreasing eigenvalues of COV and U is a unit vector of mutually orthogonal eigenvalues corresponding to each column. The normalization of COV can be expressed as:

$$COV' = U \Lambda^{\alpha} U^T \tag{23}$$

where α is a positive real number that corresponds to an unnormalized number when $\alpha = 1$ and shrinks nonlinearly for eigenvalues greater than 1, and expands nonlinearly for eigenvalues less than 1 when $\alpha < 1$. As explored by Li et al. [54], the most effective differentiation of the eigenchannels is when $\alpha = 1/2$; therefore, α was set to $1/2$.

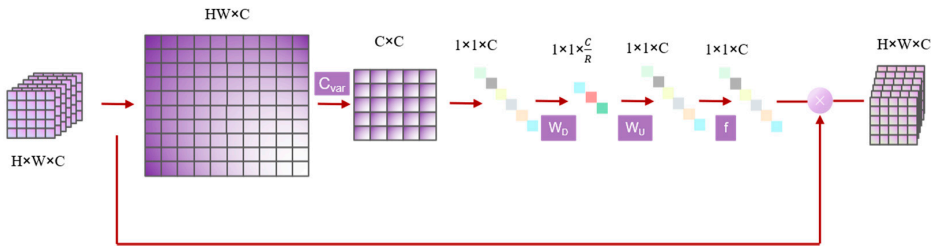


Figure 6. The process of second-order channel attention. H × W × C represents the feature map with length H, width W, and number of channels C. C_{var}, W_D, W_U, and f represent the covariance calculation, the fully connected layer for down-sampling, the fully connected layer for up-sampling and the sigmoid activation function, respectively.

At this point, we obtain the covariance matrix $COV' = [y_1, y_2, \ldots\ldots, y_n]$ of the normalized feature channels, where the channel descriptor can be obtained by pooling each line of COV':

$$Z_C = Hp(y(i)) = \frac{1}{C} \sum_1^C y(i) \tag{24}$$

where $Hp(.)$ is the pooling operation performed in the channel dimension, compared to the pooling operation in SE-Net, COV' contains the dependencies between channels and can better distinguish the feature channels. To allow the network to learn the inter-channel weights adaptively, a fully connected layer is used to perform up-sampling and down-

sampling operations on the channel descriptors Z_c. Finally, the final channel weights w are obtained by activating the sigmoid function, the process being as follows:

$$w = F(W_U(\delta(W_D(Z_c)))) \tag{25}$$

where W_U and W_D represent the up-sampling and down-sampling processes, respectively; δ denotes the Relu activation function; and F represents the sigmoid function.

2.5. Non-Local Module at the Regional Level

The conventional convolutional operation has a small perceptual field and can only use the surrounding information. In order to obtain longer dependencies, it is necessary to increase the depth of the network. However, the deepening of the network depth disrupts low-level images while gaining a larger perceptual domain, which is important for the SR task. The non-local module computes responses based on relationships between different locations, which aids the reconstruction of spatial information in the SR task [56]. However, in the SR task, the size of the feature map is too large, and the direct introduction of the non-local module is computationally intensive, while similar features in the SR task are usually within a certain range of domains; thus a region-level non-local module was adopted to obtain intra-regional dependencies and reduce the original computation. The calculation process within the non-local module is outlined in Figure 7. For an input feature map KH × KW × C, the feature map was divided into K × K regions of equal size, where the size of each region is H × W × C. In each region X, the non-local module was used to extract the dependencies, such a strategy reducing the original computational effort to $1/K^2$, and where K was set to 2 in this study. Within each region X, feature extraction was first performed using three 1 × 1 convolutions, and the number of channels was changed to c/2. We then calculated the correlation between the different positions by twice the dot product. Finally, the number of channels was changed to C by a 1 × 1 convolution layer, and the input of the module was added to the output through the residual connection.

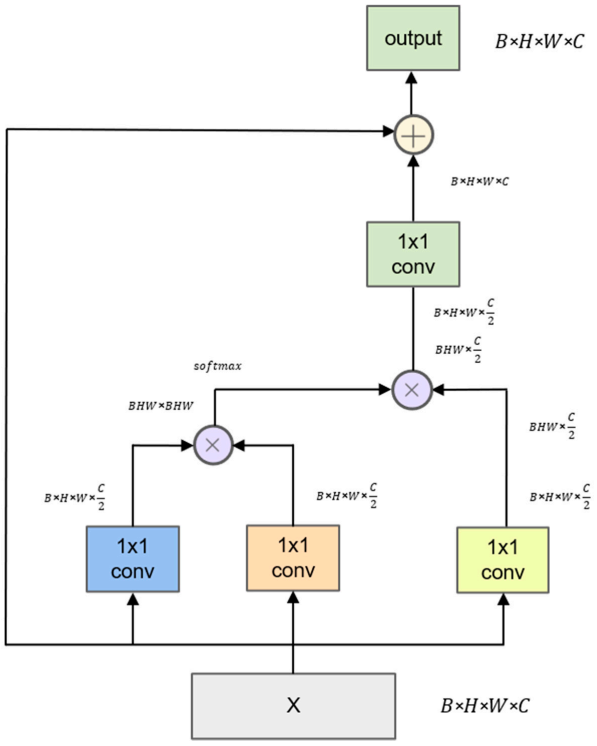


Figure 7. A non-local block. $B \times H \times W \times C$ represents the feature map with a batch size of B, length H, width W, and number of channels C. "⊗" denotes matrix multiplication, and "⊕" denotes element-wise sum.

2.6. Loss Function

To reduce artifacts, a region-aware strategy was introduced to obtain the probability distribution of the artifacts M. The artifact loss L_{art} is defined as:

$$L_{art} = \| M \cdot R_1 \|_1 \tag{26}$$

In previous work, it was demonstrated that the quality of the images could be measured by the similarity between features [57–59], thus a very deep convolutional network (VGG) [60] network was used to extract the features of the high-resolution images and super-resolution images; we then calculated the L_1 loss between the feature maps of the SR and HR images:

$$L_p = \sum_i \alpha_i \| VGG_i(I^{HR}) - VGG_i(I^{SR}) \| \tag{27}$$

where i represents the ith feature map of the VGG network before activation, and α_i represents the weight of this feature map. Referring to the SRGAN [18], the feature maps of layers 3, 4, and 5 were used, and the weights were set to $1/4$, $1/4$, $1/2$, respectively.

Unlike SRGAN, the relativistic discriminator [32] was used, and unlike the general adversarial loss, the discriminator serves to estimate the probability that the real image is relatively more realistic than the SR image, as calculated by the following expression:

$$\begin{aligned} D_R(x_r) &= \sigma(D(x_r) - \mathbb{E}_{x_f}(D(x_f))) \to 1 \\ D_R(x_f) &= \sigma(D(x_f) - \mathbb{E}_{x_r}(D(x_r))) \to 0 \end{aligned} \tag{28}$$

where D represents the discriminator network, x_r represents the HR image, x_f represents the SR image, $\mathbb{E}$ represents averaging the output of this batch, and σ represents the sigmoid activation function. Compared with the traditional discriminator, which integrates the HR image and SR image and maximizes the difference between them, the final adversarial loss consists of two parts, the adversarial loss L_G of the generator and the adversarial loss L_D of the discriminator:

$$\begin{aligned} L_G &= -\mathbb{E}_{x_r}[log(1 - D_R(x_r)] - \mathbb{E}_{x_f}[log(D_R(x_f)] \\ L_D &= -\mathbb{E}_{x_r}[log(D_R(x_r)] - \mathbb{E}_{x_f}[log(1 - D_R(x_f)] \end{aligned} \tag{29}$$

When optimizing the generator, given that $D_R(x_r)$ is fixed, $D_R(x_f)$ is made as large as possible. This strategy allows the generator to also use the discriminator to generate richer texture information; when optimizing the discriminator, it will make $D_R(x_r)$ as large as possible and $D_R(x_f)$ as small as possible.

Given that using the L_2 loss for the reconstruction of the LR images tends to result in excessive smoothing [61], L_1 loss is used to constrain the reconstructed image:

$$L_1 = E_I \| I_{HR} - I_{SR} \|_1 \tag{30}$$

E_I represents the mean value calculation.

The total losses of the network are as follows:

$$L = \lambda_1 * L_1 + \lambda_2 * L_P + \lambda_3 * L_G + \lambda_4 * L_D + \lambda_5 * L_{art} \tag{31}$$

where λ is the weight coefficient of each component. In this study, λ_1, λ_2, λ_3, λ_4, λ_5 were set to 1, 1, 0.05, 1, 1, respectively.

3. Results

In this section, the commentary is given on the following topics: the proposed dataset, comparison with other current state-of-the-art models, ablation experiments, spectral validation, and migration experiments.

3.1. Dataset

In previous studies, most datasets were produced by adding noise and blur to the down-sampled HR images to realize paired LR-HR image pairs, but the trained datasets based on this approach are often poor in terms of recovery in the real remote sensing images. To solve this problem, this study did not use publicly available datasets but built SR datasets based on four bands (red band, green band, blue band, and NIR band) GF satellite data with different spatial resolutions so that the model can learn more complex mapping relationships. The dataset consists of data from satellites GF7, GF2, GF1, and GF6, and the spatial resolution of each satellite is shown in Table 1.

Table 1. The spatial resolution of each satellite.

Satellite	PAN Spatial Resolution/m	Multi-Spectral Spatial Resolution/m
GF1	2.0	8.0
GF2	0.8	3.2
GF6	1.0	2.0
GF7	0.8	3.2

Satellite data were selected from four regions in China, that is, Beijing, Ordos, Guangzhou, and Fuzhou. In terms of time, to minimize the influences of environmental change, LR and HR images with similar times as possible were selected, of which the satellites used in each region and the times are shown in Table 2.

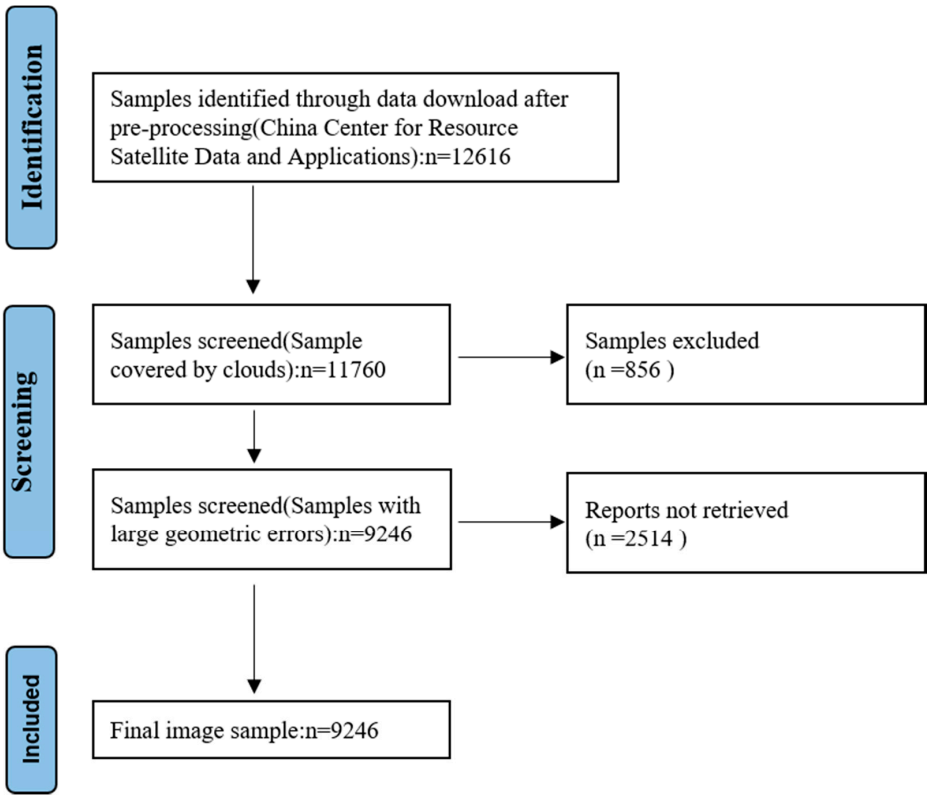
Table 2. The shooting times and satellites in the different regions. In the following table, LR and HR are relative.

City	LR Satellite	HR Satellite	LR Shooting Time	HR Shooting Time
Beijing	GF1B	GF2	1 December 2021	1 December 2021
Guangzhou	GF7	GF1	11 October 2020	11 October 2020
Fuzhou	GF1C	GF2	18 March 2022	18 March 2022
Ordos	GF6	GF7	6 July 2022	6 July 2022

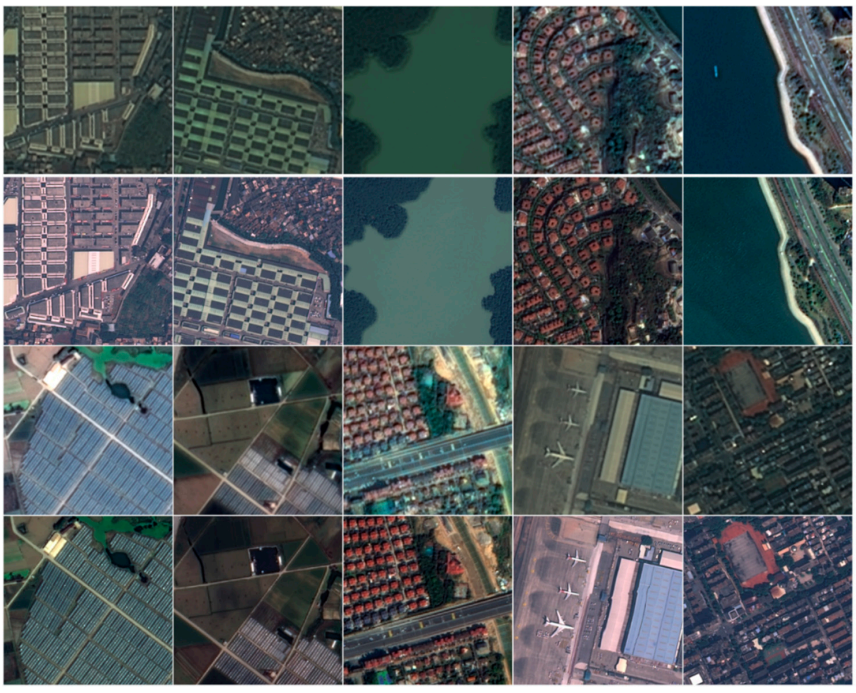
We first performed a radiometric correction based on the calibration coefficients of the GF satellites. An improved OptVM [62] is applied for band fusion. It can produce a high-resolution panchromatic image from a low-resolution multi-spectral image automatically. Afterward, the SIFT feature constraint optical flow method (SIFT-OFM) [62] is used to register images. Given that the spatial resolution of the HR image was 0.8 m, the existing model can only super-resolve the image by an integer multiple, while the spatial resolution of the LR image is 2 m; therefore, we down-sampled the HR image to 1 m by using a cubic function. Opencv2 was then used to crop the HR and LR images to 200 × 200 and 100 × 100, respectively, according to the latitude and longitude. A total of 9246 image pairs were finally generated. The images were divided randomly, with 90% serving as the training data and 10% as the validation data. The PRISMA diagram for sample identification in the study is shown in Figure 8a, and the final resulting dataset is shown in Figure 8b.

3.2. Training Details

The network is composed of two parts, the discriminator network and the generator network. The discriminator network consists of five basic blocks, and each block includes a convolutional layer responsible for feature extraction, a convolutional layer for feature map size reduction, and a BN layer where the input feature map size is 100 × 100. The generator part consists of 23 RRDCB blocks, and the number of input channels in each RRDCB block is 64; the growth of each channel in the RRDCB internal layer is 32. The generator and discriminator optimizers are based on the work of Adam [50].



(**a**)



(**b**)

Figure 8. (**a**) The PRISMA diagram for sample identification. (**b**) The proposed dataset, where the odd rows are the LR images and the even rows are the HR images.

During the training process, the input LR-HR image pairs are cropped randomly into 50 × 50 and 100 × 100 and subjected to data enhancement operations such as horizontal flip, vertical flip, and rotation. The PyTorch framework is employed to train on two Nvidia A4000 chips with a memory size of 16 GB. Further details of the experiments are given in Table 3.

Table 3. Experimental details.

Parameter	Value
Batch size	6
Traing Iter numbers	450,000
Optimization method	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.99$
Learing rate	1×10^{-4}

3.3. Evaluation Metrics

The peak signal-to-noise ratio (PSNR) [63], the structural similarity (SSIM) [64], the Frechet inception distance score (FID) [65], and the learned perceptual image patch similarity (LPIPS) [66] were selected as the evaluation metrics. The PSNR gives a measure of the degree of distortion by calculating the squared pixel-by-pixel difference between the LR and HR images; a larger PSNR value indicates that the two images are more similar. The PSNR is calculated as follows:

$$PSNR = 10 \log_{10} \frac{MAX}{MSE} \tag{32}$$

where MAX is the maximum value of pixels in the image, and the MSE is given by:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [I_{SR} - I_{HR}]^2 \tag{33}$$

SSIM gives a measure of the similarity of an image by comparing the brightness, contrast, and structure of the two images; the larger the SSIM value, the better the result for image recovery. The formula for calculating the SSIM is given by Equation (34).

$$SSIM(x, y) = l(x, y)^{\alpha} c(x, y)^{\beta} s(x, y)^{\gamma} \tag{34}$$

The PSNR and SSIM are traditional metrics for the evaluation of image quality; however, there are two problems with using only PSNR and SSIM for evaluation. First, the PSNR and SSIM do not truly reflect the quality of the images of some scenes, and higher values do not necessarily represent better quality, as is demonstrated below in the visualization of the structure. Second, the model adopted in this study is based on generative adversarial networks, and the PSNR and SSIM do not consider the relationship between the direct probability distribution of the generated samples and the real samples. Hence, the FID and LPIPS have been included in the suite of metrics to evaluate the quality of image recovery.

The FID is a metric to calculate the distance between the real image and the feature vectors of the generated image. It was shown to correlate well with the human judgment of visual quality and is most often used to evaluate the quality of samples of generative adversarial networks. FID is calculated by computing the Fréchet distance between two Gaussians fitted to feature representations of the Inception network. The higher the quality of image generation, the FID is calculated as follows:

$$FID(x, y) = \| \mu_x - \mu_y \|_2^2 + Trace(\sum_x + \sum_y - 2(\sum_x \sum_y)^{\frac{1}{2}}) \tag{35}$$

where $\| \cdot \|_2^2$ represents the square of L_2 parametrization; $Trace(.)$ represents the trace of the matrix; μ_x, μ_y are the means of the real image feature vector and the generated image, respectively; $\sum_x$, $\sum_y$ are the variances of the real image feature and the generated image feature, respectively.

The LPIPS uses a VGG network to extract features from the generated image and the real image and evaluates the similarity between the two images by measuring the square of the L_2 the parametric number between the generated image features as well as the real

image features, systematically evaluate deep features across different architectures and tasks. The value for the LPIPS is calculated as follows:

$$\sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \| \omega_l \odot (x_{hw})^l - \left(y_{hw}^l\right) \|_2^2 \tag{36}$$

where x, y represent the features of the lth feature layer of the generated image and the real image, and ω_l is a preset value for the lth feature layer weight. The smaller the LPIPS value, the more similar the generated image is to the real image.

3.4. Comparison with Existing Models

In this study, the bicubic, the super-resolution residual network (SRResNet), the enhanced deep super resolution (EDSR), the residual channel attention network (RCAN), the super-resolution generative adversarial network (SRGAN), and the enhanced super resolution generative adversarial network (ESRGAN) are used in comparison analyses with the proposed model, in which 16 residual blocks are used in the SRResNet, and the number of channels of feature maps within each residual block is 64; the EDSR uses 32 residual blocks, and the number of channels in each residual block is 256; the RCAN uses 10 RCAB groups consisting of RB basic blocks, and each RCAB group includes 20 RB blocks and the number of channels in each basic block is 64; the SRGAN has the same generator configuration as the SRResNet, the discriminator uses the VGG network where the depth of the VGG is 5, and the number of input channels to the discriminator is 64; the generator of the ESRGAN consists of 23 RRDB blocks, the number of feature channels within each basic block is 64, the number for channel growth is set to 32, each block includes five dense residual connected convolutional layers, and the discriminator is the same as for the SRGAN. These methods are retrained on our proposed dataset to achieve a fair comparison network, and the number of input channels for the discriminator is 64; moreover, to allow a fair comparison, these methods are retrained on the training set of our proposed dataset and tested on the test set.

The values of the PSNR, SSIM, LPIPS, and the FID of the model in the validation dataset after 450,000 iterations are presented in Table 4.

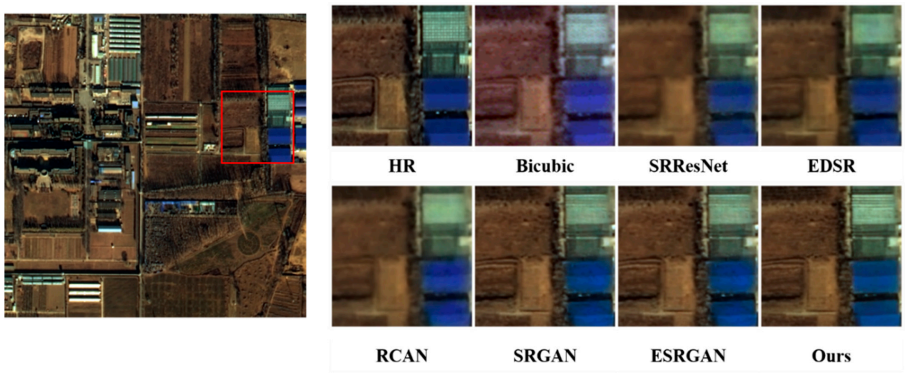
Table 4. Quantitative evaluation results for all methods in the dataset. The bold represents the best value for each indicator.

Model	**PSNR↑**	**SSIM↑**	**FID↓**	**LPIPS↓**
Bicubic	16.8941	0.5748	0.5454	71.6800
SRResNet	26.3265	0.6626	0.423	66.7573
EDSR	**27.4176**	**0.7172**	0.352	51.4357
RCAN	26.6851	0.6828	0.3995	54.0291
SRGAN	24.7729	0.5922	0.2892	20.0514
ESRGAN	24.8220	0.5906	0.2829	19.1266
Ours	26.7169	0.6890	**0.2329**	**17.9571**

As can be seen in Table 4, the EDSR achieves the best value for the PSNR and the SSIM metrics, while all the CNN models significantly outperform the GAN in terms of metrics, mainly because both the PSNR and the SSIM are computed using simple relationships between the pixel values of the image, which is similar to the definition of the loss function in the CNN networks; however, the proposed method achieves greater improvements in both metrics compared to the GAN-based methods. With respect to the PSNR and the SSIM, our model is only lower than the EDSR model by 0.7325 DB and 0.0344 DB, respectively. Our model achieves the best results for LPIPS and FID, and contrary to the previous, the GAN-based model outperforms the CNN-based model on these two metrics; moreover, the proposed method is 0.05 and 1.1695 lower than the second-ranked method on the LPIPS and FID, respectively.

In Figure 9, it can be seen that the CNN-based approach tends to generate results that are too smooth, while the GAN-based approach is able to generate more detail. Compared with other GAN-based methods, the visual quality of the present method is also the best. From the visualization results, our model has three advantages. First, we are able to produce more detail. As shown in Figure 9a, the proposed method has a clearer and sharper reduction in zebra lines, and in Figure 9b, there is a more accurate reduction in the container edges. Second, the second-order attention mechanism makes the color reproduction of the feature more accurate; as in Figure 9b, the color reproduction is closer to the original image. Third, with the help of the region-aware strategy, our image produces fewer artifacts; as can be seen in Figure 9c, the other two GAN-based methods display incorrect textures on some houses, whereas in the present method, it can basically restore the real situation of the houses correctly. Moreover, in Figure 9d, the other methods present a coarser restoration of the blue roof (see below), while our method is smoother and conforms to the HR image. In addition, from the visualization results, it can be concluded that the FID and LPIPS metrics are more in keeping with the public's perceptions compared to the PSNR and the SSIM.

Figure 9. *Cont.*



(**d**)

Figure 9. Visual comparison of some representative SR methods and the proposed model, the figure on the right is an enlargement of the area in red on the left: (**a**) road; (**b**) container; (**c**) residential buildings; (**d**) factory.

3.5. Ablation Study

To illustrate the effectiveness of the modifications, the results of several ablation experiments are considered. We gradually add region-aware strategies, second-order attention mechanisms, and region-level non-local modules to the baseline model and train them with the same configuration and test them on the validation set. The comparison data for each metric are presented in Table 5, and the entire visualization process is shown in Figure 10.

Table 5. The effect of gradual addition of different components on the performance of the model in relation to that of the baseline model. RA represents the region-aware strategy, SA represents the second-order channel attention mechanism, and NL represents the region-level non-local module. The bold represents the best value for each indicator.

Model	PSNR↑	SSIM↑	FID↓	LPIPS↓
Baseline	24.822	0.5906	0.2829	19.1266
+RA	26.2936	0.6691	0.2426	18.5648
+RA + SA	26.6243	0.6819	0.2370	**17.7798**
+RA + SA + NL	**26.7169**	**0.6890**	**0.2329**	17.9571

HR

Figure 10. Overall visual comparisons to show the effects of each component on the performance of the proposed method.

As can be seen from Table 5, the metrics of the model improve overall when adding different modules, and where the model used in this study delivers the best performance of all the metrics with the exception of the LPIPS. The three improvements are ordered in descending order of influence as region-aware strategy, second-order channel attention mechanism, and region-level non-local module, compared to the baseline model. The

model used in this work delivers the best performance in terms of the PSNR, the SSIM, the FID, and the LPIPS, the metrics being improved by 7.64%, 16.61%, 17.67%, and 6.11%, respectively. However, the LPIPS value did not improve after adding the region-level non-local module. It is considered that the NL is less appropriate for applying to the LPIPS. The LPIPS values are calculated by extracting high-level features from the network, while the non-local module is more helpful for the reconstruction of spatial information, such as edges, textures, etc., which are low-level features. In addition, high-level features are generally more concerned with global information, while low-level features are more concerned with local information; thus, there is no direct correlation between them. Our non-local module only computes in local regions of the images to reduce the computation time. The combination of the above factors causes a small decrease in the LPIPS.

As shown in Figure 10, the baseline model has a light blue artifact for the roof. After adding the region-aware strategy, the artifact disappears, and more detailed information is generated at the same time. However, the color of the building roof becomes grayish, while in the HR image, it is white. After adding the second-order channel attention mechanism, the roof color reverts to white, as observed in the HR image. Finally, by adding the non-local module, the outer contour of the building is more accurate, and at the same time, the roof becomes smoother.

3.6. Spectral Validation

In addition to visual enhancement, when further applying SR images, we need to ensure that the reflectance values are similar to those of real LR or SR satellite images. In this section, analysis was undertaken to verify that the spectral content of the SR image was similar to the real image under the different preprocessing scenarios.

The LR, HR, and SR histograms of an image from a test set are presented in Figure 11a. As can be seen, the histogram of the SR image is more similar to the HR image, which indicates that the proposed model learns the spectral information of the HR image through training. The reflectance values of 10,000 randomly selected pixels from the test data from HR and SR images are plotted in Figure 11b. From the scatter plot, an extremely strong correlation between SR images and HR images can also be found.

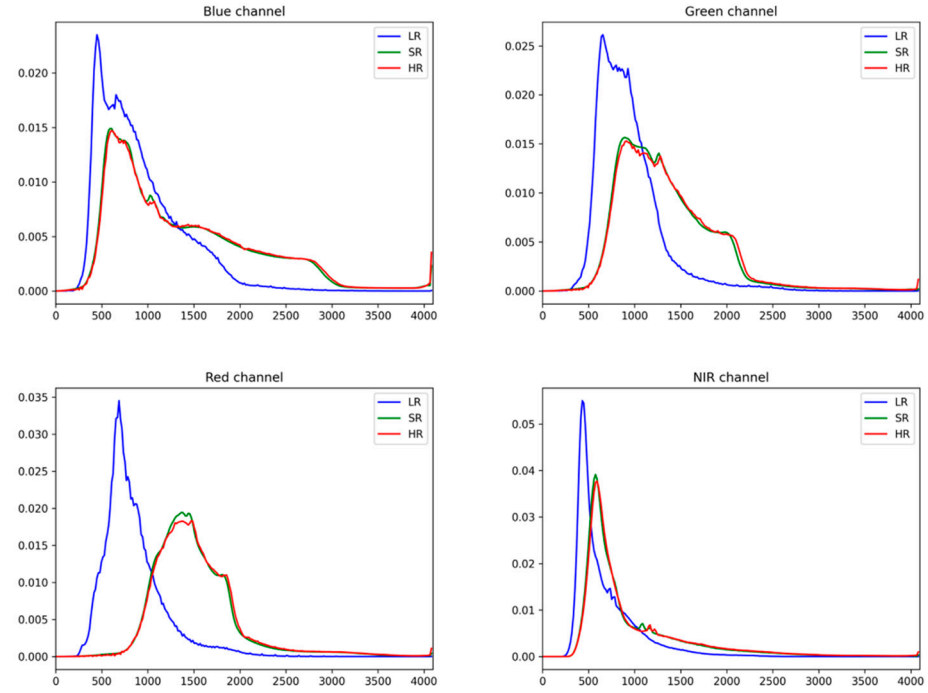
However, the spectral characteristics of both SR images and HR images are significantly different from those of LR images. We believe this is due to sensor differences between satellites. Further, to verify that the model also has the ability to retain the spectral information of the LR images. We reprocessed the images of the Guangzhou area; in addition to the preprocessing in Section 3.1, the relative radiometric correction using the histogram matching was performed on the HR images and LR images with Sentinel-2 images as the reference in the same period, and experiments were conducted on the respective test sets. As shown in Figure 12, after relative radiation corrected LR, HR and SR images have similar colors. SR images also obtained by the model can also retain the spectral information of the LR images well as a result of the additional preprocessing. Further, the reflectance values showed a strong correlation. The above experimental model can maintain the LR spectral information.

3.7. Migration Experiments

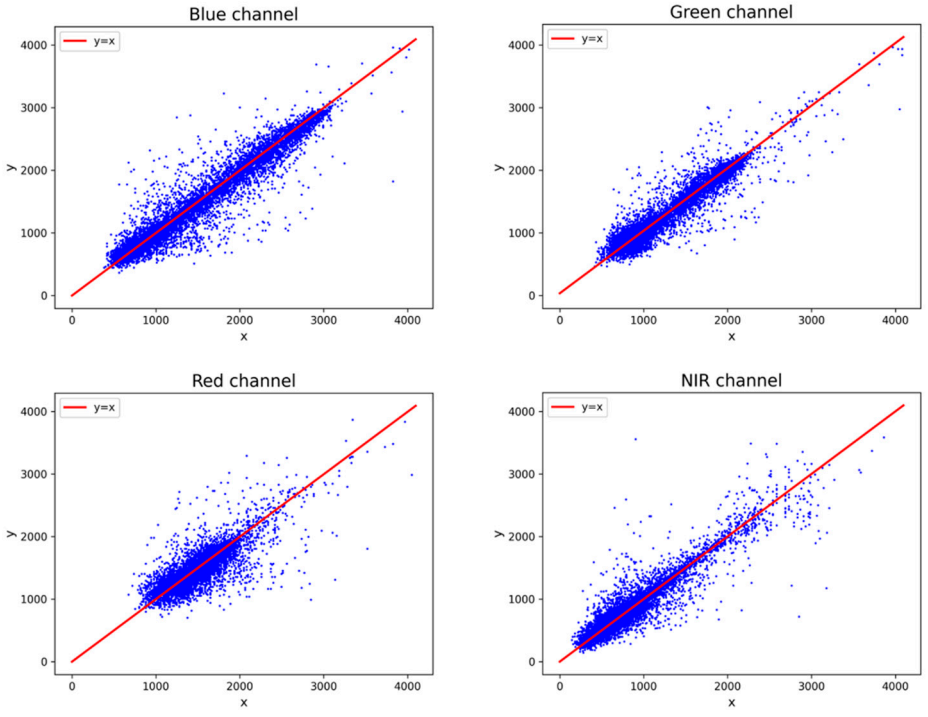
In this section, the proposed model is used to enhance the spatial resolution of the GF1 satellites from 2 m to 1 m. The datasets originate from the GF1B satellite image of Beijing on 21 October 2021 and the GF1D satellite image of Hanzhong, Shaanxi Province, on 6 July 2022. The migration experiments for the two-view remote sensing images verify the robustness of the model in time as well as in area.

The left column in Figure 13 represents the original whole-view image as well as the whole-view image after SR, where the red-rectangle regions and the zoom-in can be further visualized in terms of the local details within each area. The SR results reveal that the proposed method performs well in terms of visual quality. From the visualization results it can be seen that the edges of the image after SR are enhanced, and the information detail

is enriched, thus confirming the excellent visual performance of the model in processing real-world data.



(**a**)



(**b**)

Figure 11. Validation of the preservation of the spectral properties on our dataset: (**a**) histogram of different bands of LR, HR, SR, where the *X*-axis represents the reflectance value and the *Y*-axis represents the frequency of that reflectance; (**b**) pixel reflectance values of HR, SR scatter plot of reflectance of SR. The x-axis represents the HR reflectance, and the y-axis represents the SR reflectance.

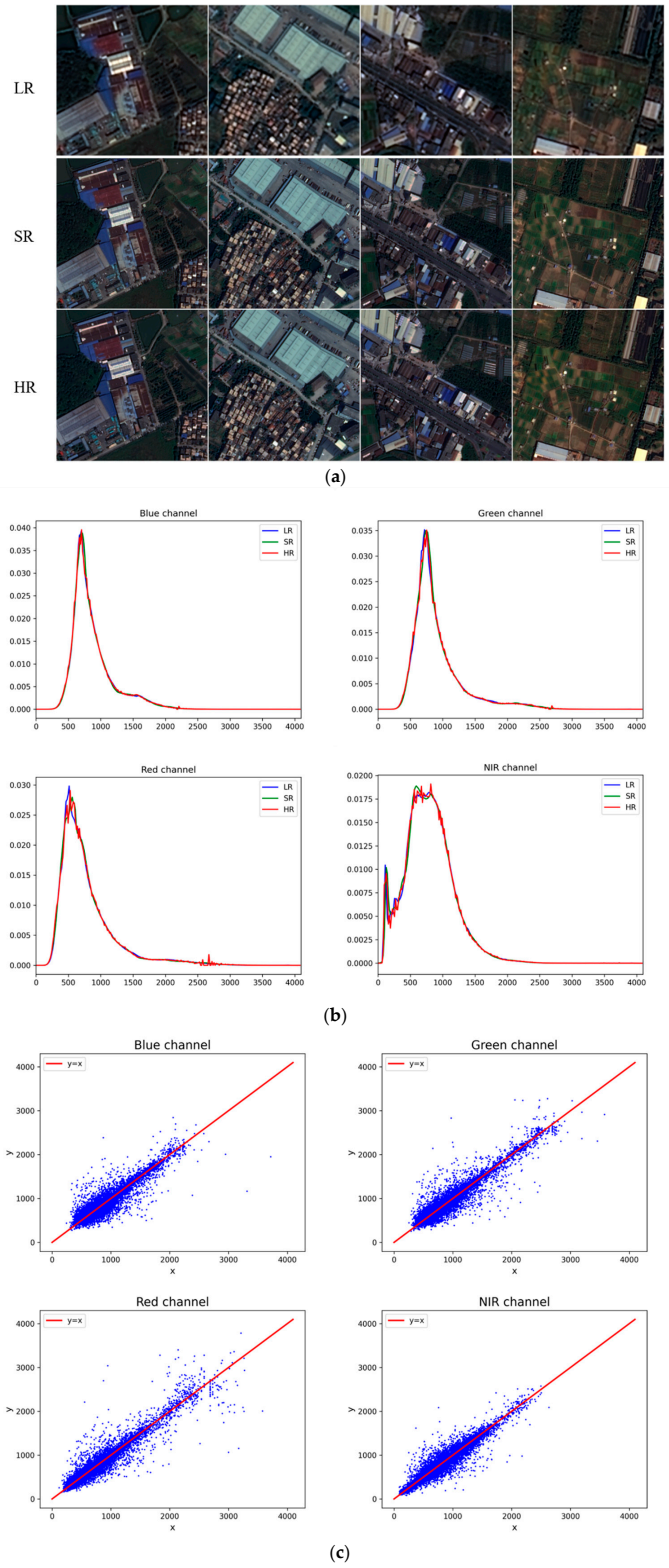
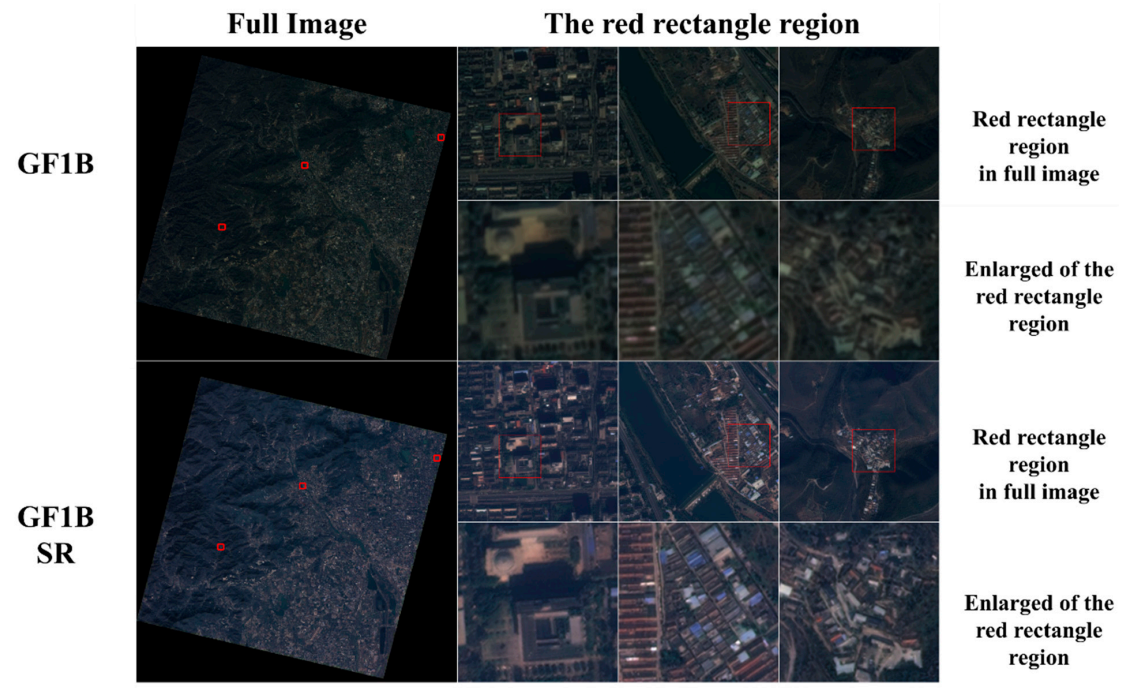
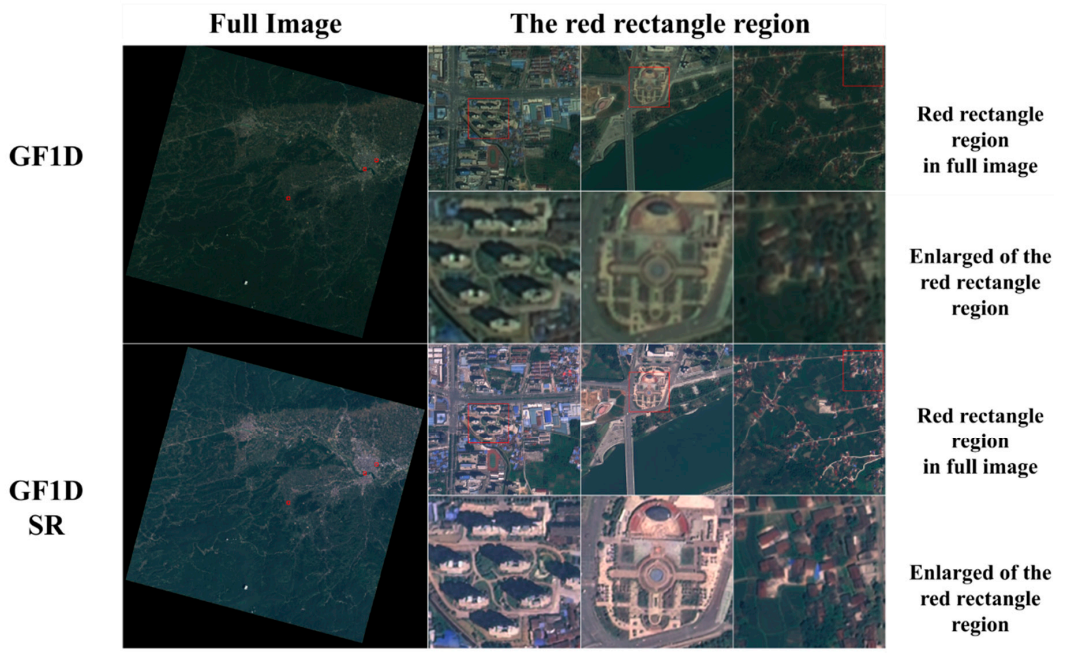


Figure 12. Validation of the preservation of the spectral properties on our dataset after relative radiometric correction: (**a**) the LR and HR images after relative radiation corrected, and SR results for the retrained model; (**b**) histogram of different bands of LR, HR, SR, where the *X*-axis represents the reflectance value, and the *Y*-axis represents the frequency of that reflectance; (**c**) pixel reflectance values of HR, SR scatter plot of reflectance of SR. The x-axis represents the HR reflectance, and the y-axis represents the SR reflectance.



(**a**)



(**b**)

Figure 13. Results of the model migration experiments in different regions and times (**a**) original images of GF1B satellite in Beijing, China, and SR results; (**b**) original images of GF1D satellite in Hanzhong, Shanxi, China, and SR results.

4. Discussion

In this study, we constructed an SR dataset of GF satellite images at various spatial resolutions to simulate a real degradation process. Previously reported models were improved, and related experiments were performed. The results in Sections 3.4 and 3.5 demonstrated that the proposed model exhibited good performance on the validation dataset. Moreover, the model outperformed all GAN-based models with respect to both the evaluation metrics and the visual aspects. In addition, we discussed the experimental results in combination with theoretical analysis.

(1) Impact of region-aware strategy: The region-aware strategy locates artifacts by taking the variance of the residual map as the basis while using the EMA. Theoretically, the variance of the artifacts on the SR and the residuals of the HR images should be larger. The smaller variance values indicated that the SR images have an identical deviation in pixel values compared to the HR images, and this only causes the wrong color to be displayed. Therefore, the small variance values in certain regions should not be judged as artifacts. The experimental results are also consistent with the aforementioned conjecture. As shown in Table 4, the new strategy can significantly improve the performance of the model, and the visualization results from Figure 9 illustrate that the artifacts can be reduced in both detail-rich regions and smooth regions, thus confirming the effectiveness of the strategy.

(2) Impact of the second-order channel attention mechanism: The second-order channel attention provides more a priori information through normalization of the covariance matrix of the feature map, allowing the network to adjust the channel weights adaptively. Its main impact is a more accurate representation of the color. As can be seen in Table 4, the method can effectively improve the performance indicators. The visualization results of Figures 9 and 10 show that this mechanism has an important role in the accurate restoration of the color of the image.

(3) Impact of the region-level non-local module: The region-level non-local module can calculate the dependency of the features that make full use of similar features in the neighboring region. It can be seen in Table 4 that the method improves the metrics other than the LPIPS. From inspection of the visualization results in Figures 9 and 10, it can be seen that this module can help in the restoration of feature contours.

(4) Comparison with other models: Compared to CNN-based approaches such as EDSR and RCAN, our model produces richer textures through adversarial learning; compared to SRGAN and ESRGAN, our method reduces the generation of artifacts while restoring color and structural information more accurately. Furthermore, most previous studies based on real satellite data used sentinel satellites [44–46], while we demonstrate the feasibility of the SR task at the meter level resolution by using GF satellites.

(5) Limits of method: First, the error in the geometric correction of the training data is basically within five pixels. If the error is too large, it will have an impact on the performance of the model. Second, due to the difference in the solar altitude angle, some higher buildings produce huge deviations in images of different spatial resolutions, and this affects the model accuracy to some extent. Third, the SOCA module calculates the covariance matrix and the eigenvalue decomposition process with high time complexity, which leads to a slow calculation speed when the input image or the number of input bands is too large.

5. Conclusions

A GAN network, which achieves good performance in data processing of the GF satellite images, has been developed. Specifically, a realistic degradation-based HR dataset using GF satellite data at different resolutions has been realized. In terms of the model, artifact generation is constrained through a region-aware strategy and the addition of a second-order channel attention mechanism at the end of the RRDCB, which adaptively learns the channel features. Finally, a region-level non-local module is added at the beginning and end of the network to take advantage of the similarity of intra-region features. Comparative experiments demonstrated that the proposed method enriches the texture details of the reconstructed images of the proposed dataset and that the results for the reconstructed images are superior to previous methods with respect to the FID and LPIPS. Moreover, we can keep the spectral information very well. The migration experiments for the GF1 satellites further demonstrate the effectiveness of the proposed method in different times and regions. Due to the improvement of spatial resolution, the proposed SR method is designed as a preprocessing step for small object detection, fine land cover classification, high time–frequency change detection, etc. In future work, two issues will be addressed. First, we intend to perform 4× SR work based on real remote sensing satellites, such as

using the GF satellite and the Quick bird satellite images. Second, there is a problem of ground LR-HR mismatch in real scenes due to the use of different satellite shooting angles. Correction modules will be added to the network with the aim of solving this problem.

Author Contributions: Conceptualization, J.Z. and F.C.; Data curation, W.Y. and J.Y.; Formal analysis, J.Z.; Funding acquisition, Y.M. and W.Y.; Investigation, J.Z.; Methodology, J.Z. and F.C.; Project administration, Y.M. and F.C.; Resources, Y.M. and J.Y.; Software, J.Z. and S.Z.; Supervision, Y.M. and F.C.; Validation, J.Z. and E.S.; Visualization, J.Z. and E.S.; Writing—original draft, J.Z.; Writing—review and editing, J.Z., Y.M. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Nature Science Foundation of China (Grant No. 42201063), the Key Research and Development Program of Hainan Province (ZDYF2021SHFZ260), and Hainan Provincial Natural Science Foundation of China (322QN345).

Data Availability Statement: The data of experimental images used to support the findings of this research are available from the corresponding author upon reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

HR	High resolution
LR	Low resolution
SR	Super resolution
SA-GAN	Second order attention generator adversarial network
GF	GaoFen
CNN	Convolutional neural network
SRResNet	Super-resolution residual network
EDSR	Enhanced deep super resolution
RCAN	Residual channel attention network
GAN	Generating adversarial network
SRGAN	Super resolution generative adversarial network
Esrgan	Enchanced super resolution generative adversarial network
RRDCB	Residual in residual dense channel attention block
VGG	Very deep convolutional networks
SOCA	Second order channel attention
RA	Region aware
NL	Non-local
PSNR	Peak signal-to-noise ratio
SSIM	Structural similarity
FID	Frechet inception distance score
LPIPS	Learned perceptual image patch similarity

References

1. Manandhar, R.; Odeh, I.O.A.; Ancev, T. Improving the Accuracy of Land Use and Land Cover Classification of Landsat Data Using Post-Classification Enhancement. *Remote Sens.* **2009**, *1*, 330–344. [CrossRef]
2. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep Convolutional Neural Networks for the Scene Classification of High-Resolution Remote Sensing Imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
3. Xu, Y.; Wu, L.; Xie, Z.; Chen, Z. Building Extraction in Very High Resolution Remote Sensing Imagery Using Deep Learning and Guided Filters. *Remote Sens.* **2018**, *10*, 144. [CrossRef]
4. Peng, D.; Zhang, Y.; Guan, H. End-to-End Change Detection for High Resolution Satellite Images Using Improved UNet++. *Remote Sens.* **2019**, *11*, 1382. [CrossRef]
5. Gu, J.; Sun, X.; Zhang, Y.; Fu, K.; Wang, L. Deep Residual Squeeze and Excitation Network for Remote Sensing Image Super-Resolution. *Remote Sens.* **2019**, *11*, 1817. [CrossRef]
6. Rukundo, O.; Cao, H. Nearest neighbor value interpolation. *arXiv* **2012**, arXiv:1211.1768.

7. Zhang, X.-G. A New Kind of Super-Resolution Reconstruction Algorithm Based on the ICM and the Bilinear Interpolation. In Proceedings of the 2008 International Seminar on Future BioMedical Information Engineering, Wuhan, China, 18 December 2008; pp. 183–186.
8. Zhang, X.-G. A New Kind of Super-Resolution Reconstruction Algorithm Based on the ICM and the Bicubic Interpolation. In Proceedings of the 2008 International Symposium on Intelligent Information Technology Application Workshops, Washington, DC, USA, 21–22 December 2008; pp. 817–820.
9. Rasti, P.; Demirel, H.; Anbarjafari, G. Image Resolution Enhancement by Using Interpolation Followed by Iterative Back Projection. In Proceedings of the 2013 21st Signal Processing and Communications Applications Conference (SIU), Haspolat, Turkey, 24–26 April 2013; pp. 1–4.
10. Wheeler, F.W.; Hoctor, R.T.; Barrett, E.B. Super-Resolution Image Synthesis Using Projections onto Convex Sets in the Frequency Domain. In Proceedings of the Computational Imaging III, San Jose, CA, USA, 17 January 2005; pp. 479–490.
11. Chantas, G.K.; Galatsanos, N.P.; Woods, N.A. Super-resolution based on fast registration and maximum a posteriori reconstruction. *IEEE Trans. Image Process.* **2007**, *16*, 1821–1830. [CrossRef]
12. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [CrossRef]
13. Dong, C.; Loy, C.C.; Tang, X. Accelerating the Super-Resolution Convolutional Neural Network. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 391–407.
14. 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.
15. 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.
16. Tai, Y.; Yang, J.; Liu, X. Image Super-Resolution via Deep Recursive Residual Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 3147–3155.
17. 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, 21–26 July 2017; pp. 136–144.
18. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z. 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.
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
20. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 448–456.
21. 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; pp. 286–301.
22. Guo, Y.; Chen, J.; Wang, J.; Chen, Q.; Cao, J.; Deng, Z.; Xu, Y.; Tan, M. Closed-Loop Matters: Dual Regression Networks for Single Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5407–5416.
23. Lee, J.; Jin, K.H. Local Texture Estimator for Implicit Representation Function. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1929–1938.
24. Wang, L.; Dong, X.; Wang, Y.; Ying, X.; Lin, Z.; An, W.; Guo, Y. Exploring Sparsity in Image Super-Resolution for Efficient Inference. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4917–4926.
25. 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 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
26. Yang, F.; Yang, H.; Fu, J.; Lu, H.; Guo, B. Learning Texture Transformer Network for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 5791–5800.
27. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
28. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 1833–1844.
29. 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, Nashville, TN, USA, 20–25 June 2021; pp. 10012–10022.
30. Chen, X.; Wang, X.; Zhou, J.; Dong, C. Activating More Pixels in Image Super-Resolution Transformer. *arXiv* **2022**, arXiv:2205.04437.

31. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
32. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Change Loy, 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.
33. Jolicoeur-Martineau, A. The relativistic discriminator: A key element missing from standard GAN. *arXiv* **2018**, arXiv:1807.00734.
34. Wang, X.; Yu, K.; Dong, C.; Loy, C.C. Recovering Realistic Texture in Image Super-Resolution by Deep Spatial Feature Transform. In Proceedings of the Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 606–615.
35. Wang, X.; Xie, L.; Dong, C.; Shan, Y. Real-esrgan: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 1905–1914.
36. Zhang, K.; Zuo, W.; Zhang, L. Learning a Single Convolutional Super-Resolution Network for Multiple Degradations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3262–3271.
37. Xu, Y.-S.; Tseng, S.-Y.R.; Tseng, Y.; Kuo, H.-K.; Tsai, Y.-M. Unified Dynamic Convolutional Network for Super-Resolution with Variational Degradations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12496–12505.
38. Gu, J.; Lu, H.; Zuo, W.; Dong, C. Blind Super-Resolution with Iterative Kernel Correction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1604–1613.
39. Wang, L.; Wang, Y.; Dong, X.; Xu, Q.; Yang, J.; An, W.; Guo, Y. Unsupervised Degradation Representation Learning for Blind Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 10581–10590.
40. Luo, Z.; Huang, Y.; Li, S.; Wang, L.; Tan, T. Learning the Degradation Distribution for Blind Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 6063–6072.
41. Yue, Z.; Zhao, Q.; Xie, J.; Zhang, L.; Meng, D.; Wong, K.-Y.K. Blind Image Super-Resolution with Elaborate Degradation Modeling on Noise and Kernel. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2128–2138.
42. 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]
43. Galar, M.; Sesma, R.; Ayala, C.; Albizua, L.; Aranda, C. Super-resolution of sentinel-2 images using convolutional neural networks and real ground truth data. *Remote Sens.* **2020**, *12*, 2941. [CrossRef]
44. Salgueiro Romero, L.; Marcello, J.; Vilaplana, V. Super-resolution of sentinel-2 imagery using generative adversarial networks. *Remote Sens.* **2020**, *12*, 2424. [CrossRef]
45. Zabalza, M.; Bernardini, A. Super-Resolution of Sentinel-2 Images Using a Spectral Attention Mechanism. *Remote Sens.* **2022**, *14*, 2890. [CrossRef]
46. Karwowska, K.; Wierzbicki, D. Improving Spatial Resolution of Satellite Imagery Using Generative Adversarial Networks and Window Functions. *Remote Sens.* **2022**, *14*, 6285. [CrossRef]
47. Zhang, Z.; Tian, Y.; Li, J.; Xu, Y. Unsupervised Remote Sensing Image Super-Resolution Guided by Visible Images. *Remote Sens.* **2022**, *14*, 1513. [CrossRef]
48. Guo, M.; Zhang, Z.; Liu, H.; Huang, Y. NDSRGAN: A Novel Dense Generative Adversarial Network for Real Aerial Imagery Super-Resolution Reconstruction. *Remote Sens.* **2022**, *14*, 1574. [CrossRef]
49. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Deep Image Prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9446–9454.
50. 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, 18–24 June 2022; pp. 5657–5666.
51. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
52. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning. In Proceedings of the Thirty-first AAAI conference on artificial intelligence, San Francisco, CA, USA, 4–9 February 2017.
53. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
54. Li, P.; Xie, J.; Wang, Q.; Zuo, W. Is Second-Order Information Helpful for Large-Scale Visual Recognition? In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–27 October 2017; pp. 2070–2078.
55. Dai, T.; Cai, J.; Zhang, Y.; Xia, S.-T.; Zhang, L. Second-order attention network for single image super-resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11065–11074.
56. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-Local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.

57. 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, Amsterdam, The Netherlands, 8–16 October 2016; pp. 694–711.
58. Bruna, J.; Sprechmann, P.; LeCun, Y. Super-resolution with deep convolutional sufficient statistics. *arXiv* **2015**, arXiv:1511.05666.
59. Dosovitskiy, A.; Brox, T. Generating images with perceptual similarity metrics based on deep networks. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Barcelona, Spain; 5–10 December 2016; pp. 658–666.
60. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
61. Zhao, H.; Gallo, O.; Frosio, I.; Kautz, J. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging* **2016**, *3*, 47–57. [CrossRef]
62. Ma, Y.; Chen, F.; Liu, J.; He, Y.; Duan, J.; Li, X. An automatic procedure for early disaster change mapping based on optical remote sensing. *Remote Sensing* **2016**, *8*, 272. [CrossRef]
63. Korhonen, J.; You, J. Peak Signal-to-Noise Ratio Revisited: Is simple beautiful? In Proceedings of the 2012 Fourth International Workshop on Quality of Multimedia Experience, Melbourne, Australia, 5–7 July 2012; pp. 37–38.
64. 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]
65. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6629–6640.
66. 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.