

Article

Super-Resolution Image Reconstruction Method between Sentinel-2 and Gaofen-2 Based on Cascaded Generative Adversarial Networks

Xinyu Wang ^{1,2} , Zurui Ao ³, Runhao Li ^{2,4}, Yingchun Fu ^{2,*}, Yufei Xue ² and Yunxin Ge ³

¹ Guangdong Center for Marine Development Research, Guangzhou 510220, China

² School of Geography, South China Normal University, Guangzhou 510631, China

³ Beidou Research Institute, South China Normal University, Guangzhou 528225, China

⁴ China Water Resources Pearl River Planning, Surveying and Designing Co., Ltd., Guangzhou 510610, China

* Correspondence: fuyyc@m.scnu.edu.cn

Abstract: Due to the multi-scale and spectral features of remote sensing images compared to natural images, there are significant challenges in super-resolution reconstruction (SR) tasks. Networks trained on simulated data often exhibit poor reconstruction performance on real low-resolution (LR) images. Additionally, compared to natural images, remote sensing imagery involves fewer high-frequency components in network construction. To address the above issues, we introduce a new high-resolution dataset GF_Sen based on GaoFen-2 and Sentinel-2 images and propose a cascaded network CSWGAN combined with spatial-frequency features. Firstly, based on the proposed self-attention GAN (SGAN) and wavelet-based GAN (WGAN) in this study, the CSWGAN combines the strengths of both networks. It not only models long-range dependencies and better utilizes global feature information, but also extracts frequency content differences between different images, enhancing the learning of high-frequency information. Experiments have shown that the networks trained based on the GF_Sen can achieve better performance than those trained on simulated data. The reconstructed images from the CSWGAN demonstrate improvements in the PSNR and SSIM by 4.375 and 4.877, respectively, compared to the relatively optimal performance of the ESRGAN. The CSWGAN can reflect the reconstruction advantages of a high-frequency scene and provides a working foundation for fine-scale applications in remote sensing.

Keywords: super-resolution; remote sensing; cross-sensor; generative adversarial network; cascade structure



Citation: Wang, X.; Ao, Z.; Li, R.; Fu, Y.; Xue, Y.; Ge, Y. Super-Resolution Image Reconstruction Method between Sentinel-2 and Gaofen-2 Based on Cascaded Generative Adversarial Networks. *Appl. Sci.* **2024**, *14*, 5013. <https://doi.org/10.3390/app14125013>

Academic Editor: Atsushi Mase

Received: 30 March 2024

Revised: 1 June 2024

Accepted: 4 June 2024

Published: 8 June 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing data from different sensors vary in temporal and spatial resolutions. These sensors images have been widely used in many fields, such as Sentinel-2, which plays an important role in many applications [1], including monitoring vegetation [2,3], urban extraction [4], and the construction of hydrological models [5], while GaoFen-2 focuses on fine land cover classification and feature extraction [6]. The sensor design must balance capturing spatial details and addressing repeated coverage concerns [7]. It is particularly challenging to obtain high-quality images in areas with cloudy or rainy conditions. The integration of two sensors offers vital advantages in constructing dense image time series with a high spatial resolution, essential for the long-term and detailed monitoring of surface processes. Therefore, the development of super-resolution reconstruction (SR) methods to improve the resolution of Sentinel-2 images to generate synthetic high-resolution time series datasets has attracted extensive attention.

The traditional image SR algorithm is mainly based on interpolation or traditional machine learning [8,9]. Due to the reliance of Bicubic methods on local neighborhood information, unsatisfactory results have been achieved [8,10]. The emergence of deep

learning has shown promising results in SR. Convolution neural networks (CNNs) have made significant progress in image SR [11–17]. SRCNN produces clearer reconstructed images and significantly improves the reconstruction speed [18]. ESPCN accelerates feature learning and reduces network complexity by directly passing feature information in the LR space [19]. However, with the increasing demand for reconstruction results, the depth of the network is also increasing, leading to the problem of vanishing gradients. The residual structure has gradually played a role in SR tasks. It allows CNNs to model more deeply without suffering from vanishing/exploding gradients [11]. Subsequently, some heavyweight models have been proposed [13,16,20–22]. However, due to excessive focus on minimizing the mean square reconstruction error, the reconstruction results are “overly smooth” [23–25].

Generative adversarial networks (GANs) have been introduced to image SR by leveraging adversarial learning between the generator and discriminator to generate high-quality results. For example, SRGAN demonstrates the ability of GANs in inverse imaging problems, which combines multiple loss functions to ensure that the generated results have both good perceptual and objective quality [26]. ESRGAN improves reconstruction by addressing noise susceptibility and enhancing residual blocks with residual dense blocks [27]. In the field of remote sensing, inspired by the natural image reconstruction method, research based on the characteristics of remote sensing images has also been carried out. Meng et al. had addressed the challenges of large-scale remote sensing images by combining dense sampling and residual learning methods to enhance the effectiveness of feature extraction [28]. Despite the benefits of convolutions in modeling image dependencies, their ability to learn long-term dependencies is limited. To address this problem, the self-attention mechanism is introduced to carefully coordinate the details of each position of the image with the details of the distant part of the image [29]. Combining the self-attention mechanism with GANs allows for the extraction of global features without introducing excessive parameters [30].

In recent years, due to its advantage of a self-attention mechanism, the transformer model has become an essential choice in the field of natural language processing. However, the complexity of the transformer model has posed a significant challenge to its performance. To address this issue, an Efficient Super-Resolution Transformer (ESRT) is proposed. The ESRT aims to enhance the ability of SR networks to capture long-range contextual dependencies while significantly reducing the GPU memory cost [31]. However, due to the rich land coverage and complex land use scenes contained in remote sensing images within a limited pixel size, the loss of high-frequency detail information is more severe. It is difficult to effectively reconstruct the lost high-frequency information by focusing only on spatial-domain information. Restoring high-frequency information through power spectrum features in the frequency domain appears simpler and more effective. Wavelet transform combines temporal and frequency resolutions, adapting window sizes and facilitating detail feature extraction. It can process high-frequency information such as edges and details while reducing computational memory [32–34]. In DWSR, residual networks and wavelet transforms exploit the residual sparsity of wavelet coefficients. Different wavelet subbands, providing complementary structural information in different directions, predict the HR image [35]. WRAN [36] uses a multi-kernel convolutional to adaptively aggregate features of the receptive fields from different-sized receptive fields of wavelet coefficients. Nevertheless, the multi-band image information of different sensors is complex and degraded and there are few studies on the frequency-domain multi-scale reconstruction of image details. Combining wavelet transform and the residual structure of GANs ensures network gradients and extracts deeper features [37]. In addition, loss constraint in the wavelet domain retains the advantages of wavelet transform in high-frequency information reconstruction and avoids excessive smoothing of reconstructed images [38].

However, most of the existing algorithms extract the spatial- and frequency-domain features using a single network. Deepening or widening the network improves algorithm performance, it also increases computational complexity, and at the same time, it is difficult

for these methods to effectively couple spatial and frequency information. Therefore, there is a need to develop a new network model that can maintain the advantages of extracting spatial information while enhancing the ability to extract frequency-domain information.

So far, although many SR methods based on remote sensing have been developed, it has been challenging to acquire paired high-resolution and low-resolution images for training, leading to a lack of corresponding “high-resolution and low-resolution” real remote sensing image pairs. Additionally, most models are trained on natural image datasets, like DIV2K, UC_Merced [39], and WHU-RS19 [40], which results in significant performance degradation when applied to real remote sensing image SR tasks [41]. To better address the degradation information between HR and LR remote sensing images, this study focuses on developing a benchmark paired remote sensing image dataset considering the following four critical features: (1) different sensors, (2) diversity of regions, (3) consistency of time, and (4) challenging scenarios. Based on the four aforementioned aspects, we propose a new benchmark dataset GF_Sen, formed by images obtained from GaoFen-2 and Sentinel-2.

The main objectives of this study are as follows: (1) To propose a cascaded spatial frequency-domain generative adversarial network CSWGAN, specifically for cross-sensor remote sensing images reconstruction. We first proposed the SGAN and WGAN, which integrate the advantages of a self-attention mechanism and wavelet transform, enriching the global feature information and high-frequency details in the reconstructed image. The CSWGAN combines the advantages of two networks through a cascaded structure, making its reconstruction results closer to HR images; (2) to propose a new benchmark paired remote sensing dataset GF_Sen for SR tasks, to evaluate the performance of several widely used deep learning models (SRCNN, ESPCN, SRGAN, ESRGAN, etc.). This dataset aims to provide a fair and comprehensive comparison between different models.

2. Materials and Methods

In this section, we will first give a brief introduction to the two main methods used in the CSWGAN. Then, we will introduce the proposed method including the design of the network architectures and loss functions.

2.1. Self-Attention Mechanism

The self-attention module through 1×1 convolutions calculates the correlation matrix with the C channel feature map, which represents the spatial correlation between any two positions in the input characteristic graph, as shown in Figure 1. Each location is calculated and updated by the weighted sum of all other locations. Among them, the weight value is determined by the learning dependency between the two positions. The formula for calculating the extent of different locations in each area of the reconstructed image is as follows:

$$\beta_{j,i} = \exp(s_{ij}) / \sum_{i=1}^N \exp(s_{ij}), \text{ where } s_{ij} = f(x_i)^T g(x_j) \quad (1)$$

where $\beta_{j,i}$ indicates the extent to which the model attends to the i th location when synthesizing the j th region. N is the number of feature locations of features. The output of the attention layer o is defined as

$$o_j = v \left(\sum_{i=1}^N \beta_{j,i} h(x_i) \right), h(x_i) = W_h x_i, v(x_i) = W_v x_i \quad (2)$$

In the above formulation, $W_h \in \mathbb{R}^{\bar{C} \times C}$ and $W_v \in \mathbb{R}^{C \times \bar{C}}$ are the learned weight matrices, which are implemented as 1×1 convolutions. Among them, $\bar{C} = C/8$. In addition, we further multiply the output of the attention layer by a scale parameter and add back the input feature map. Therefore, the final output is given by

$$y_i = \gamma o_i + x_i \quad (3)$$

where γ is a learnable scalar and it is initialized as 0, which allows the network to first rely on the information in the local neighborhood, and then gradually learn to assign more weight to the non-local evidence.

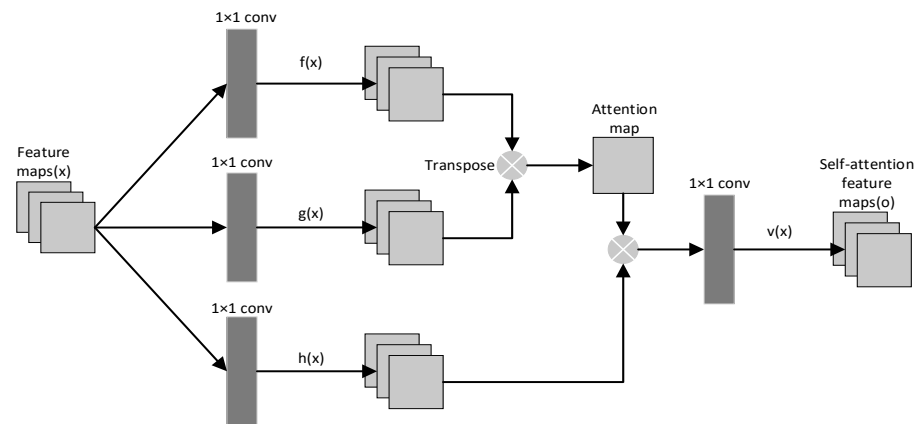


Figure 1. The self-attention mechanism for CSWGAN.

2.2. Wavelet Prediction and Reconstruction

In this paper, the wavelet transform method is used to decompose the characteristic image obtained by the residual module into a sequence of wavelet coefficients of the same size. The wavelet prediction net can be split into $N_w = 4n$ parallel independent subnets, where n is the decomposition level. The corresponding wavelet coefficients are obtained from the input feature extraction block output of each subnet. It is assumed that the characteristic size of wavelet coefficients is the same as that of the LR input.

The reconstruction net is used to transform the wavelet images of the total size $N_w \times 3 \times h \times w$ into the original image space of the size $3 \times (r \times h) \times (r \times w)$. It is represented by $r \times r$ size filter, composed of a deconvolution layer with a step size of r , and N_e is the channel size of the last layer of the feature extraction block. Let $C = (c_1, c_2, \dots, c_{N_w})$ and $\hat{C} = (\hat{c}_1, \hat{c}_2, \dots, \hat{c}_{N_w})$ denote the ground-truth and inferred wavelet coefficients, respectively. The overall process can be defined as

$$\begin{aligned} \hat{y} &= \phi(\hat{C}) = \phi\{\{\hat{c}_1, \hat{c}_2, \dots, \hat{c}_{N_w}\}\} = \phi\{(\varphi_1(\hat{z}), \varphi_2(\hat{z}), \dots, \varphi_{N_w}(\hat{z}))\} \\ &= \phi\{(\phi_1(\psi(x)), \phi_2(\psi(x)), \dots, \phi_{N_w}(\psi(x)))\} \end{aligned} \quad (4)$$

where $\psi : \mathbb{R}^{3 \times h \times w} \rightarrow \mathbb{R}^{N_e \times h \times w}$, $\varphi_i : \mathbb{R}^{N_e \times h \times w} \rightarrow \mathbb{R}^{3 \times h \times w}$, $i = 1, 2, \dots, N_w$,

$$\phi : \mathbb{R}^{N_w \times 3 \times h \times w} \rightarrow \mathbb{R}^{3 \times (r \times h) \times (r \times w)}.$$

2.3. A Cascaded Spatial Frequency-Domain Generative Adversarial Network for SR: CSWGAN

This paper proposed the CSWGAN, which was a cascade network of a self-attention-based GAN (SGAN) and wavelet-based GAN (WGAN), as shown in Figure 2.

The main idea is to design a cascade framework to implement the application of spatial and frequency-domain features to obtain high-quality SR images.

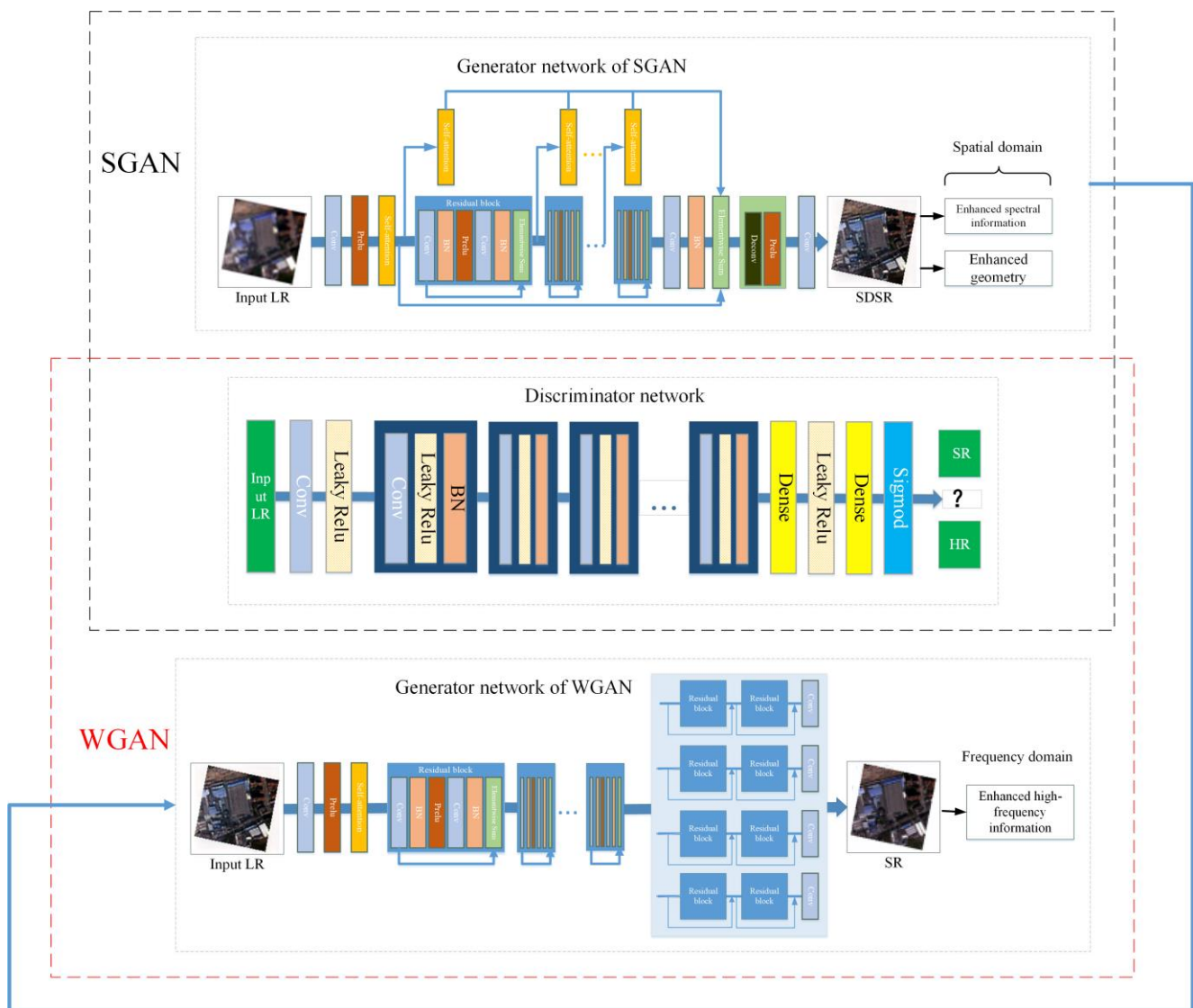


Figure 2. Network architecture of CSWGAN.

The main architecture is based on SRGAN [26], with residual blocks serving as an important component in the network construction, enabling the recovery of photo-realistic textures on public benchmarks. The activation functions used in the generator and discriminator are different, with the generator employing PReLU to allow negative values to pass through the activation function. This helps to address the vanishing gradient problem and enhances the overall performance of the network by enabling the model to learn more complex data relationships. In the discriminator, LeakyReLU activation is used to prevent the need for max-pooling throughout the network.

In order to enhance the learning of the spectral and geometric position information in remote sensing images, a self-attention branch is added outside the residual block of the backbone network. By incorporating self-attention mechanisms into feature extraction methods, dense global contextual information can be obtained, thereby establishing interdependencies between pixels. This enhances the network's utilization of global information and deep features. Then, the SR task is transformed into a wavelet coefficient prediction task, where the wavelet prediction module is tightly integrated with the wavelet reconstruction module within the GAN. By accurately predicting the corresponding wavelet coefficients, the LR image is decomposed into different scales and orientations, resulting in a high-quality HR image with rich texture details and global feature information.

The main task of the SGAN network is to learn and obtain the geometric structure information of the image, as well as low-frequency information such as the overall appearance and contours of the image. On the other hand, the WGAN is designed according to the characteristics of the frequency domain. Its main task is to learn high-frequency information such as image texture and edge. The SGAN learns spatial low-frequency features and their correlations through a self-attention mechanism. Based on the output characteristics of the SGAN, the WGAN can better distinguish noise and detail information, further optimize low-frequency and enhance high-frequency learning, thus improving the quality and fidelity. Therefore, by using the spatial-domain super-resolution (SDSR) output of the SGAN as the input to the WGAN and generating the final reconstruction output, the CSWGAN network can fully leverage the advantages of these two different networks. Moreover, training this cascade structure using multiple generators and discriminators is very time-consuming, so each adversarial system is trained separately first and then jointly in an end-to-end manner.

2.4. Loss Function

Since the CSWGAN is logically composed of two independent networks, there is no overall loss function. The loss functions of the SGAN and WGAN work in their respective networks.

2.4.1. Loss Function of SGAN

In this paper, I^{LR} and I^{HR} are used as the LR image and HR image, respectively. We construct a robust loss function to force generator G to generate an HR image similar to I^{HR} ,

$$L_{SGAN} = L_{mse} + \alpha_1 L_{adv} + \alpha_2 L_{tv} \quad (5)$$

where α_1 and α_2 represent the coefficients of L_{adv} (adversarial loss) and L_{tv} (regularization loss), with values of 1×10^{-3} and 2×10^{-8} . The network requires the reconstruction effect at the pixel level through L_{mse} , trains the discriminator D to minimize the adversarial loss, and encourages G to generate a reconstructed image close to the I^{HR} . It is well known that adversarial learning strategies can maintain the visual authenticity of the generated images; however, it is also vulnerable to noise pollution. In order to restrain the noise and keep the image smooth, TV regularization loss is introduced in this paper.

$$L_{mse} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y} \right)^2 \quad (6)$$

$$L_{adv} = \sum_{n=1}^N -\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \quad (7)$$

$$L_{tv} = \frac{1}{r^2WH} \sum_{x=1}^{rW} \sum_{y=1}^{rH} \left\| \nabla G_{\theta_G}(I^{LR})_{x,y} \right\| \quad (8)$$

where N denotes the number of images, θ_G denotes a set of model parameters in G , θ_D denotes a set of model parameters in D , $G_{\theta_G}(I^{LR})$ denotes the reconstructed image, $D_{\theta_D}(G_{\theta_G}(I^{LR}))$ denotes the probability that the $G_{\theta_G}(I^{LR})$ is a real HR image, and $\nabla G_{\theta_G}(I^{LR})$ represents the gradient of the output of G .

2.4.2. Loss Function of WGAN

Different to the spatial-domain-based network that directly learns the mapping relationship between HR and LR pixel values, the WGAN aims to learn the relationship between the wavelet coefficients of frequency-domain characteristics between HR and LR

images. The WGAN constrains the network through two losses in the spatial and frequency domain of the image.

$$L_{\text{WGAN}} = L_{\text{wavelet}} + \beta L_{\text{texture}} + L_{\text{full-image}} \quad (9)$$

where β is the balance parameter, which is 0.1. The WGAN takes the MSE loss function $L_{\text{full-image}}$ as the loss constraint of the full image to make the reconstruction result achieve a balance in smoothness and texture details; it can hardly obtain high-frequency texture details. Therefore, in order to pay attention to local texture information and prevent high-frequency wavelet coefficients from converging to 0, the WGAN adopts L_{wavelet} and L_{texture} avoids the above problems, respectively.

$$L_{\text{wavelet}}(\hat{C}, C) = \sum_{i=1}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2 = \lambda_1 \|\hat{c}_1 - c_1\|_F^2 + \sum_{i=2}^{N_w} \lambda_i \|\hat{c}_i - c_i\|_F^2 \quad (10)$$

$$L_{\text{texture}} = \sum_{i=k}^{N_w} \gamma_i \max(\alpha \|c_i\|_F^2 + \varepsilon - \|\hat{c}_i\|_F^2, 0) \quad (11)$$

where λ_i is the balance parameter, c_i and $\hat{c}_i$ are the ground-truth wavelet coefficients and the inferred wavelet coefficient, respectively. k indicates the start index of the wavelet coefficients to be penalized for taking small values, γ_i represents the balance weight, α and ε are slack values. More attention can be paid to the local texture with large weight assigned to the high-frequency coefficient. At the same time, it can capture the global topology information.

2.5. Quantitative Evaluation Indices

The evaluation of super-resolution results mainly includes subjective evaluation and objective evaluation. We used various image quality evaluation indicators, including peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [42], standard deviation (SD), information entropy (IE) [43], root mean square error (RMSE), and spectral angle (SAM) [42], as shown in Table 1. These indicators not only measure the fidelity and differences between real HR images and SR images on the ground, but they can also better explain the amount of information contained in the image and evaluate many subtle differences from the perspective of spectrum. SD and IE are non-reference image quality evaluators. Additionally, three experts with three years of experience in this field compared the naturalness of SR images.

Table 1. The equations of quantitative evaluation indices.

Indicator	Formula	Remarks
PSNR	$20 \log_{10} \frac{255}{\text{RMSE}}$	/
SSIM	$\sum_j^K \left(\frac{(2\bar{X}\bar{Y} + c_1)(2\sigma_{XY} + c_2)}{(\bar{X}^2 + \bar{Y}^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)} \right)_j$	X and Y correspond to the reconstructed image and the original image, c_1 and c_2 constants should be set to (K_1L) and (K_2L) , respectively, where K_1 and K_2 are values close to 0, and L represents the image dynamic range.
RMSE	$\sqrt{\frac{1}{K \times N} \sum_j^K \sum_i^N (X_i^j - Y_i^j)^2}$	N is the total number of pixels in each image, K is the number of bands.
SAM	$\frac{1}{N} \sum_i^N \arccos \left(\frac{X_i \cdot Y_i}{\ X_i\ \ Y_i\ } \right)$	/
SD	$\sqrt{\frac{1}{N} \sum_i^N (F(i) - \bar{u})^2}$	$F(i)$ represents the pixel value, $\bar{u}$ represents the mean pixel value.
IE	$-\sum_i^N P(i) \log_2 P(i)$	$P(i)$ represents the probability of the pixel value.

3. Experiments

3.1. Datasets and Preprocessing

In this section, we describe our benchmark dataset GF_Sen, which consists of imagery obtained from Sentinel-2 and GaoFen-2. Simultaneously, three publicly available datasets, UC_Merced, WHU-RS19, and USC-SIPI, are introduced.

3.1.1. Real-World Multi-Sensor LR-HR Dataset: GF_Sen

To enhance the network's generalization on sensor and spatiotemporal dimensions, this paper constructs a set of paired real-world multi-sensor LR-HR data, in which Sentinel-2 data with 10 m spatial resolution are used as the LR image, and GaoFen-2 data with 4 m spatial resolution are used as the HR images. This dataset is referred to as GF_Sen, and the details of the dataset are shown in Table 2. The HR images are downsampled to generate LR images as simulated data.

Table 2. The information of used data.

Location	Sensor	Filename
Guangzhou	Sentinel-2	L1C_T49QGF_A003347_20191002T030635
Guangzhou	GaoFen-2	GF2_PMS1_E113.2_N23.3_20191227_L1A0004507303
Guangzhou	Sentinel-2	L1C_T49QGF_A003347_20191002T030635
Guangzhou	GaoFen-2	GF2_PMS2_E113.4_N23.1_20191227_L1A0004505416
Shenzhen	Sentinel-2	L1C_T49QGF_A003347_20171027T030705
Shenzhen	GaoFen-2	GF2_PMS1_E113.8_N22.8_20171227_L1A0002883454
Shenzhen	Sentinel-2	L1C_T49QGF_A003347_20171027T030705
Shenzhen	GaoFen-2	GF2_PMS2_E114.0_N22.6_20171227_L1A0002883537
Dongguan	Sentinel-2	L1C_T49QHF_A003347_20171027T030705
Dongguan	GaoFen-2	GF2_PMS2_E114.1_N22.9_20171227_L1A0002883531
Huizhou	Sentinel-2	L1C_T50QKL_A007994_20170102T025445
Huizhou	GaoFen-2	GF2_PMS1_E114.5_N23.1_20161128_L1A0001994661

To build a paired dataset for SR, the two sensors need to image the same ground targets at the same positions in the electromagnetic spectrum. In this study, we finally chose three bands of blue, green, and red. The study area includes four different cities in Guangdong Province, each with diverse land cover and ecological types. Some sample images of the dataset are shown in Figure 3.

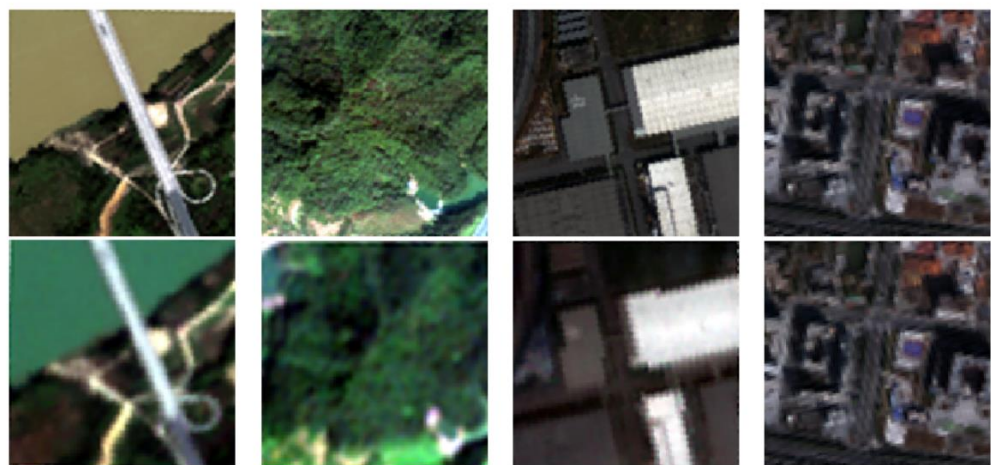


Figure 3. LR-HR image-pair samples in GF_Sen dataset. The first row is HR images from GaoFen-2, while the second row is the corresponding LR images from Sentinel-2.

Although searching different satellite data within a small time window can reduce variations caused by atmospheric conditions and environmental changes, only a limited number of images meet such strict criteria. Therefore, we expanded the time window to 3 months. After observation, there were basically only a few active objects (such as ships and vehicles) exhibiting differences in this period. We searched the data of the study area (as shown in Figure 4), selected the images with less than 5% of the total cloud cover, and finally deleted the image data samples with clouds. After resampling the GaoFen-2 data using a bilinear sampling method to a 5 m resolution, we registered the Sentinel-2 data with reference to the GaoFen-2 data.

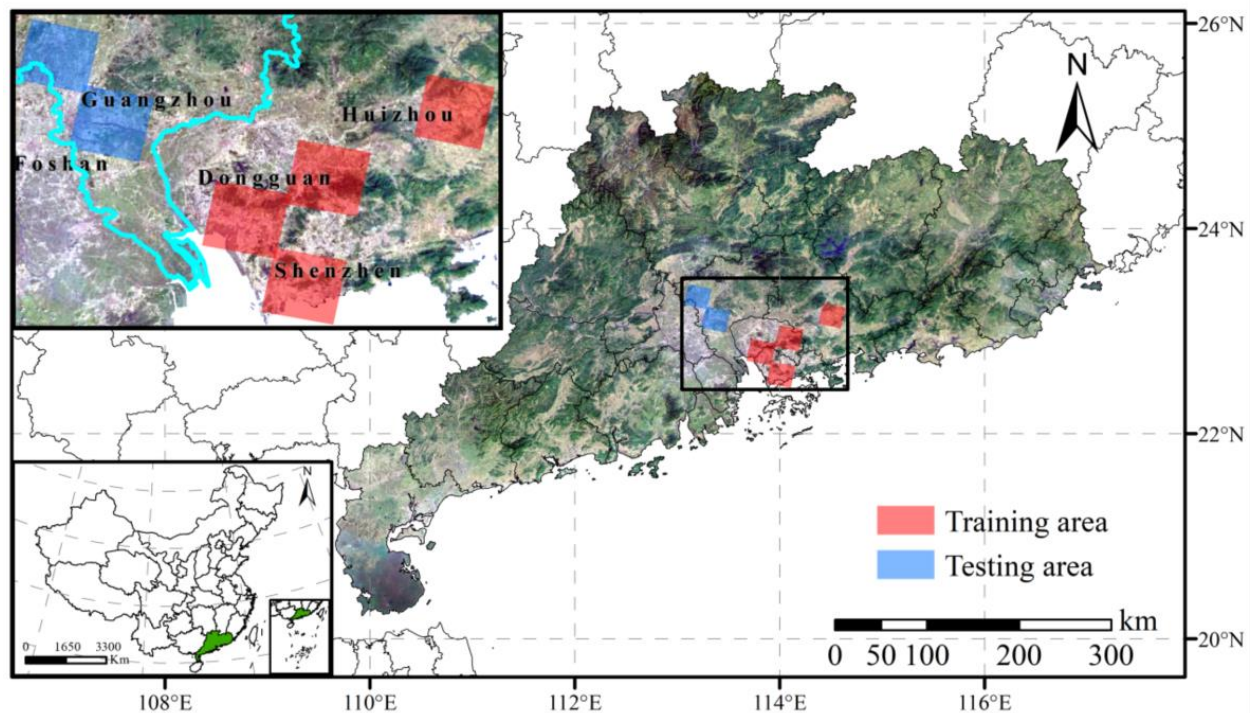


Figure 4. Spatial distribution of Sentinel-2 and GaoFen-2.

Following the preprocessing of the images mentioned above, the HR images were cropped to a size of 128×128 , while the LR images were cropped to a size of 64×64 . The GF_Sen dataset was finally created, consisting of 5263 images for the training set, 2255 images for the validation set, and 3930 images for the test set.

3.1.2. UC_Merced

To pre-train the model, this paper utilized the UC_Merced dataset, with a total of 2100 images selected from aerial orthophotos, each with a size of 256×256 pixels. The original dataset has 21 categories, and each category consists of 100 images. In this paper, each scene is randomly selected as training and validation sets with a ratio of 7:3.

3.1.3. WHU-RS19

To evaluate the network's performance on different scene reconstruction results, this paper also selected WHU-RS19 as the test set for the comparative experiments. The WHU-RS19 dataset consists of 950 aerial images. The size of each image in the dataset is 600×600 pixels. The WHU-RS19 dataset contains 19 categories of land use types, and each category consists of 50 images.

3.1.4. USC-SIPI

To assess the reconstruction performance of the model on large-scale remote sensing images, this study selected images from the “aerials” category in the dataset for testing. The database is divided into volumes based on the basic character of the pictures. Images in each volume are of various sizes such as 256×256 pixels, 512×512 pixels, or 1024×1024 pixels.

3.2. Training Details

In the training process, the batch size is set to 16, and the training process is divided into two parts. In the first part, the SGAN was trained to obtain the SDSR, the learning rate was initialized to 10^{-3} , and halved every 200 epochs. A total of 1600 epochs were trained. In the second part, we used the SDSR as the input LR image again to train the WGAN, where the decomposition level was set to 1, the learning rate was initialized to 10^{-4} , with a decay rate of 0.0005.

4. Results

4.1. CSWGAN Evaluation

We extensively evaluated the SR performance from different networks trained on both the simulation dataset and GF_Sen dataset.

From Table 3, it can be seen that the network accuracy is generally higher when trained on the GF_Sen dataset compared to the simulation dataset, and the CSWGAN achieved the best results. Specifically, on the GF_Sen dataset, compared with the relatively excellent ESRGAN, the CSWGAN was competitive due to the combination of spatial–frequency-domain features, with an average improvement of 4.877 in the PSNR value and 0.181 in the SSIM value. This proved that our approach of restoring high-frequency information in images using a cascaded network structure is highly effective. By taking a comprehensive consideration of the results in Table 3 and Figure 5, we can find that the CSWGAN can produce acceptable super-resolution results compared to classical networks, especially in terms of texture structure. Finally, the CSWGAN combined the advantages of both the SGAN and WGAN, and the edge contour of the recovered image was clearer, the aliasing effect was effectively improved, and the overall tone of the image was closer to the HR image.

In addition, due to the characteristics of remote sensing images, compared with small-sized image blocks, they are more likely to cover a larger range of scenes. Therefore, this study compared the reconstruction results of several classic networks in the GF_Sen and USC-SIPI datasets, under larger and even larger scales of scenes, as shown in Figures 6 and 7. It can be seen that the reconstruction results of the CSWGAN outperform other methods in both high-frequency information and visual perception.

Table 3. Quality evaluation index results of different SR methods.

Method	Trained on Simulated Dataset				Trained on GF_Sen Dataset			
	PSNR	SSIM	SAM	RMSE	PSNR	SSIM	SAM	RMSE
Bicubic					18.863	0.593	0.301	30.039
SRCNN	17.975	0.506	0.284	28.506	18.191	0.512	0.316	30.553
ESPCN	18.006	0.612	0.302	29.864	18.266	0.582	0.303	29.864
ESRT	19.092	0.461	0.219	24.484	20.287	0.669	0.191	24.484
SRGAN	18.654	0.611	0.385	36.675	19.217	0.639	0.372	36.168
ESRGAN	18.975	0.621	0.286	28.948	19.458	0.647	0.338	33.060
CSWGAN	23.350	0.853	0.192	17.781	24.335	0.828	0.178	15.356

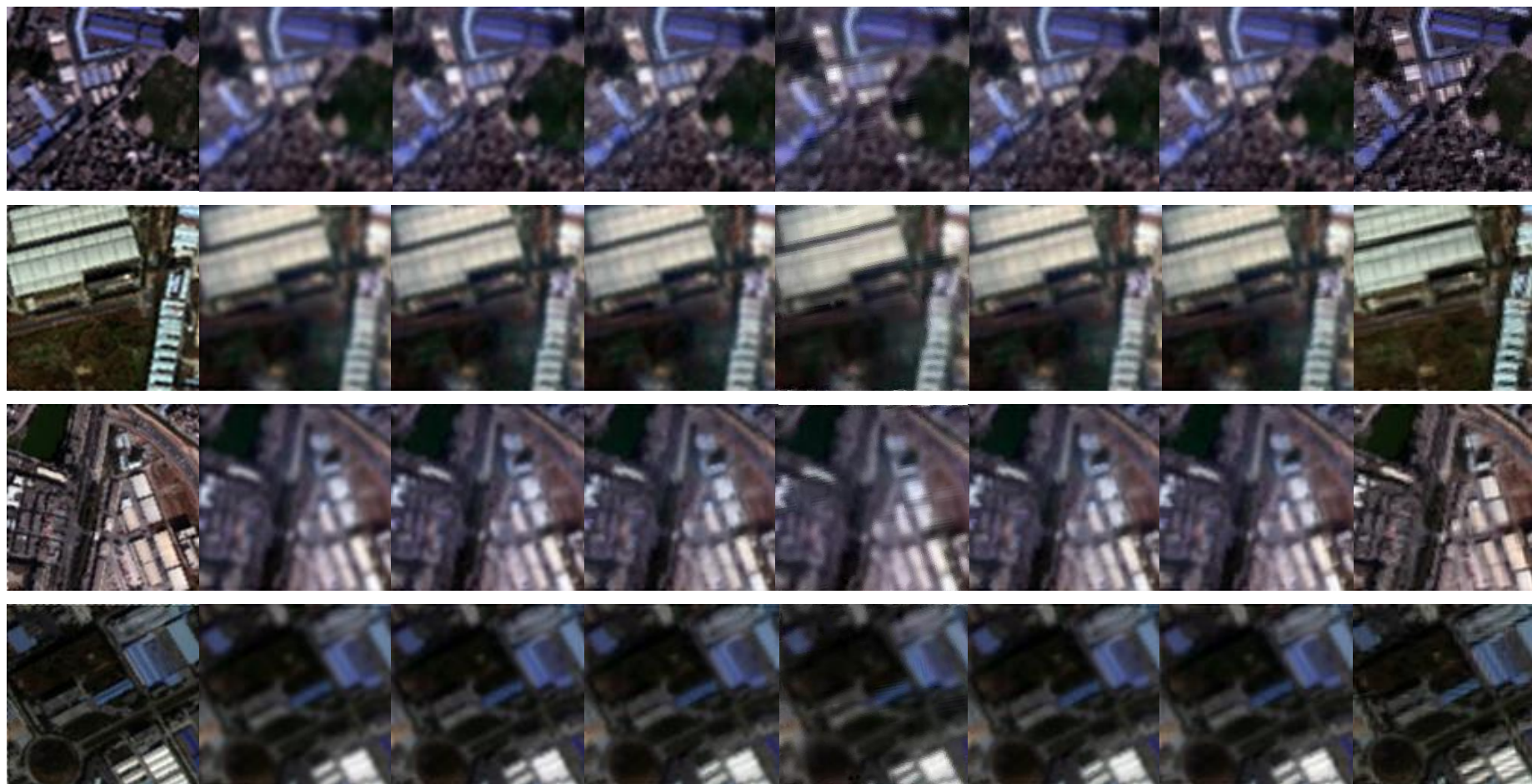


Figure 5. Cont.

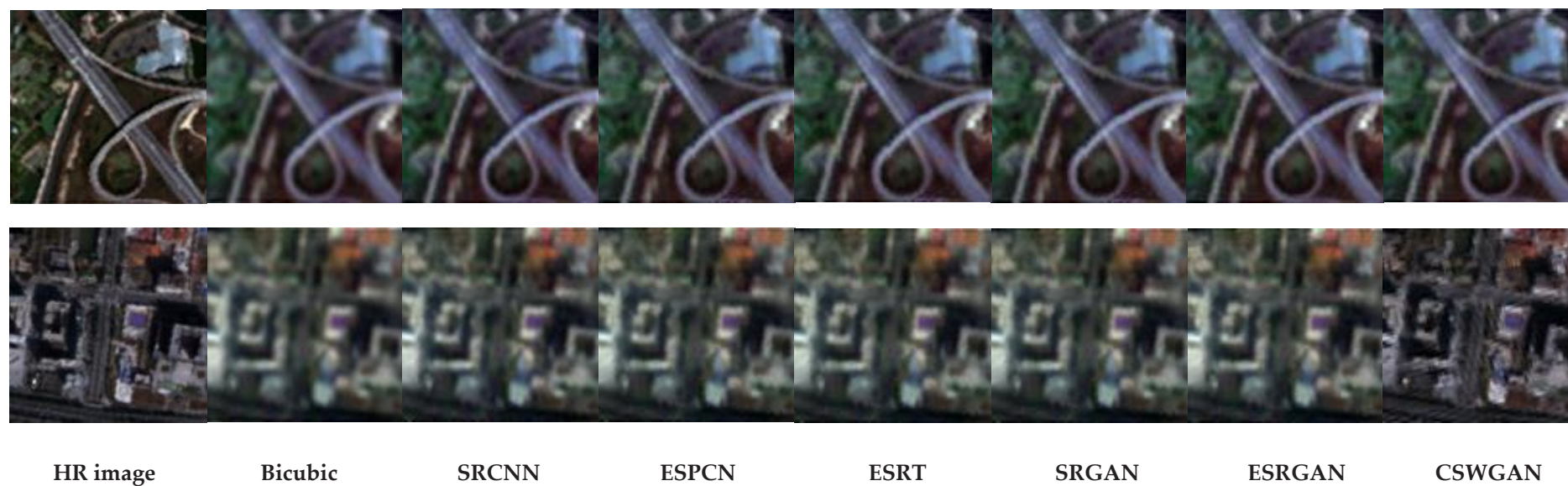


Figure 5. Reconstruction results of different methods.

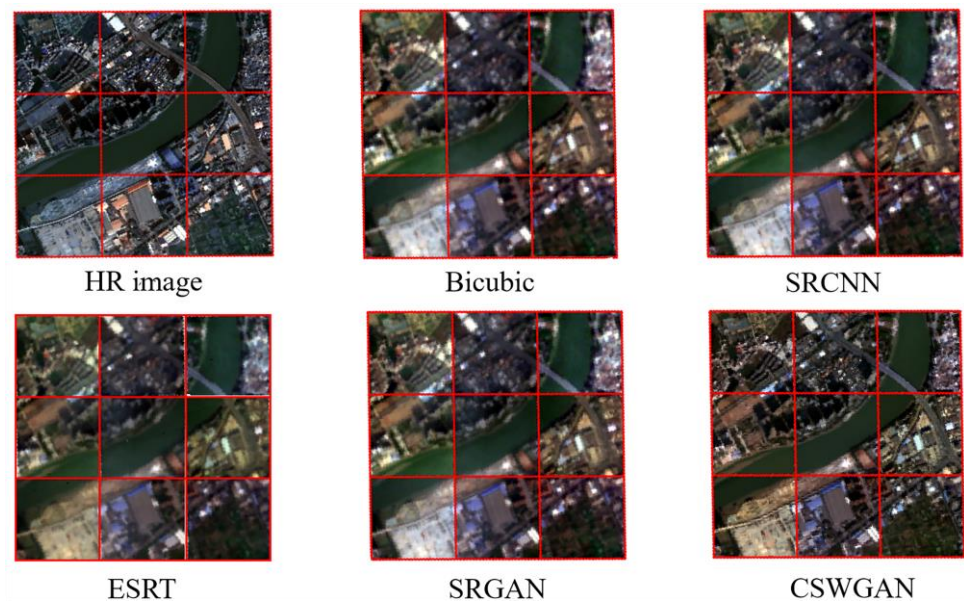


Figure 6. Reconstruction results of GF_Sen.

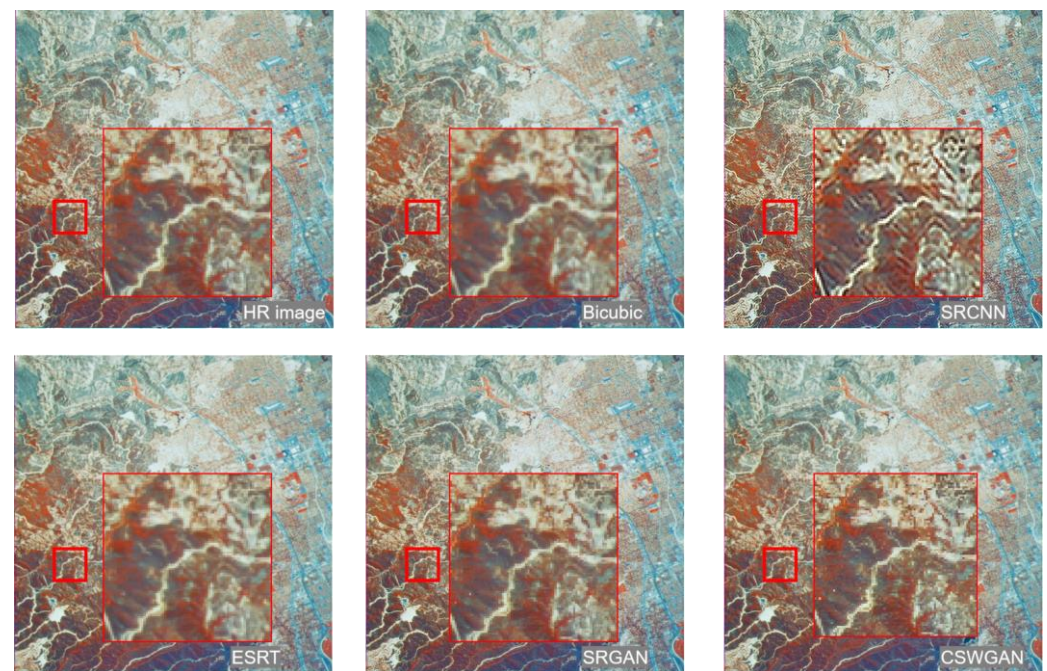


Figure 7. Reconstruction results of USC-SIPI.

4.2. Comparison of CSWGAN SR on Multi-Sensor Scenes

We evaluated the generative performance of the SR network, which demonstrated the ability to fit the manifold of HR images. In order to explore the practical significance of SR, this study reconstructed different types of scenes on the WHU-RS19 test set using the CSWGAN and comparative methods, demonstrating the differences in reconstruction effects.

The reconstruction results of different scenes are shown in Table 4. It can be seen that the reconstruction effect of the CSWGAN is generally better than that of other networks. Specifically, for scenes with rich high-frequency information such as bridges, commercial buildings, residential buildings, and industrial areas, the reconstruction results of the CSWGAN can reflect the reconstruction advantages of high-frequency information. However, for scenes with less high-frequency information and mostly low-frequency infor-

mation, such as forest and meadow, the use of a lightweight network can save time and obtain better reconstruction results.

Table 4. Comparison of reconstruction accuracy of different scenes. A–G represent Bicubic, SRCNN, ESPCN, ESRT, SRGAN, ESRGAN, and CSWGAN, respectively, and the bold value is the optimal result.

Type	psnr_A	ssim_A	psnr_B	ssim_B	psnr_C	ssim_C	psnr_D	ssim_D	psnr_E	ssim_E	psnr_F	ssim_F	psnr_G	ssim_G
Airport	32.054	0.925	32.623	0.944	32.908	0.941	25.559	0.769	30.309	0.907	29.928	0.859	32.393	0.947
Beach	36.257	0.987	37.429	0.992	34.400	0.977	42.324	0.969	36.165	0.976	37.341	0.922	45.142	0.993
Bridge	34.183	0.942	34.789	0.942	34.276	0.944	30.548	0.875	34.74	0.957	31.463	0.927	34.859	0.958
Commercial area	27.621	0.887	28.946	0.919	28.862	0.909	22.692	0.698	26.112	0.870	27.193	0.845	28.967	0.920
Forest	32.113	0.884	33.362	0.923	31.908	0.901	26.599	0.666	30.180	0.855	30.408	0.868	32.555	0.910
Industrial area	30.489	0.915	31.379	0.938	31.214	0.932	24.374	0.738	28.488	0.896	19.350	0.661	31.620	0.941
Meadow	42.375	0.955	40.131	0.968	39.402	0.963	34.17	0.84	41.052	0.946	40.512	0.918	41.316	0.963
Mountain area	27.368	0.827	28.490	0.877	27.758	0.862	23.529	0.613	26.438	0.819	27.443	0.845	28.152	0.870
Park	33.142	0.904	33.987	0.929	33.613	0.925	26.274	0.73	31.697	0.895	31.336	0.839	34.084	0.932
Parking	30.993	0.947	32.261	0.965	31.756	0.962	24.27	0.801	27.869	0.922	29.084	0.917	32.140	0.958
Pond	46.854	0.984	42.743	0.981	39.981	0.955	29.934	0.848	47.010	0.984	39.921	0.967	44.387	0.981
Port	32.247	0.940	33.063	0.954	32.002	0.944	24.906	0.808	30.290	0.927	31.742	0.969	33.151	0.954
Train station	28.084	0.873	30.210	0.926	29.450	0.917	23.902	0.669	25.926	0.847	27.002	0.898	30.274	0.930
Residential area	27.756	0.907	29.345	0.934	29.161	0.924	22.434	0.725	25.677	0.883	28.249	0.863	29.395	0.935
River	32.266	0.898	33.105	0.921	32.550	0.918	27.148	0.746	30.587	0.881	30.077	0.898	32.608	0.929
Viaduct	28.330	0.896	29.887	0.930	29.955	0.924	23.497	0.705	30.255	0.926	26.361	0.874	30.477	0.935

Furthermore, the CSWGAN is trained on real high- and low-resolution data, while the WHU dataset uses synthesized pairs of high- and low-resolution images. Therefore, the CSWGAN performed poorly in some scenes when facing the WHU dataset. However, in most scenarios, it achieved better performance than other models, indicating that the CSWGAN can demonstrate good universality for both simulated and real datasets.

4.3. Ablation Study

4.3.1. Evaluate the Performance of SGAN in Spatial Domain

We analyzed the objective evaluation metrics of reconstructed images to evaluate the contribution of the SGAN reconstructed images based on spatial information. Additionally, to assess the improvement in the network training on the GF_Sen dataset, this paper trained different networks based on the simulated dataset and GF_Sen, and then tested them on the real LR images to obtain different image quality evaluation indicators.

Table 5 compares the test accuracy of the networks trained based on the two training datasets. It was found that the accuracy of most networks trained with the GF_Sen was improved, as shown in the fact that the PSNR and SSIM values were generally higher than the network accuracy trained on the simulated datasets. The PSNR or SSIM cannot fully represent the reconstruction quality. In this paper, the SD and IE were added based on the statistical characteristics of images, which aim to evaluate the network effect through gray-level distribution and the amount of information the image has, as shown in Table 5.

Table 5. Comparison accuracy of different SR methods in different test sets.

Method	Trained on Simulated Dataset				Trained on GF_Sen Dataset			
	PSNR	SSIM	SD	IE	PSNR	SSIM	SD	IE
Bicubic					18.863	0.593	61.411	6.699
SRCNN	17.975	0.506	53.999	5.757	18.191	0.512	59.795	6.281
ESPCN	18.006	0.612	59.257	6.331	18.266	0.582	60.257	6.447
ESRT	19.092	0.461	63.2142	6.3867	19.287	0.669	64.0794	6.162
SRGAN	18.654	0.611	58.713	6.416	19.217	0.639	63.049	6.663
ESRGAN	18.975	0.621	62.588	6.502	19.458	0.647	63.280	6.802
SGAN	19.452	0.641	63.893	6.820	19.552	0.645	64.593	6.833

Further analysis of the evaluation results showed that SRGAN and ESRGAN were higher than Bicubic, SRCNN, and ESPCN in four indicators, which proved that the GAN can better utilize spatial information than Bicubic and CNN. At the same time, the ESRT also demonstrated advantages over the GAN in some indicators, believing that the combination of a self-attention mechanism and transformer can reconstruct images with more saturated

colors. Finally, the SGAN obtained the highest PSNR and SSIM values on the GF_Sen test set, which were 19.552 and 0.645, respectively, which proved that the combination of the self-attention mechanism and GAN can make the reconstructed images also have a more diverse range of grayscale levels and abundant information in the spatial domain.

4.3.2. Evaluate the Performance of WGAN in Frequency Domain

Due to the limited use of frequency-domain-based super-resolution reconstruction networks, it was important to verify the detailed features of the WGAN-reconstructed images from the frequency-domain perspective.

In this paper, three images with gradually increasing high-frequency information were selected for analysis, as shown in Figure 8. Because of the difficulty in recovering high-frequency information from remote sensing images, wavelet transformation was employed to extract the high-frequency information of different images, including the horizontal high frequency (cH), vertical high frequency (cV), and diagonal high frequency (cD) information. The SD and IE of the HR image and the results of each model are compared, as shown in Figure 9.

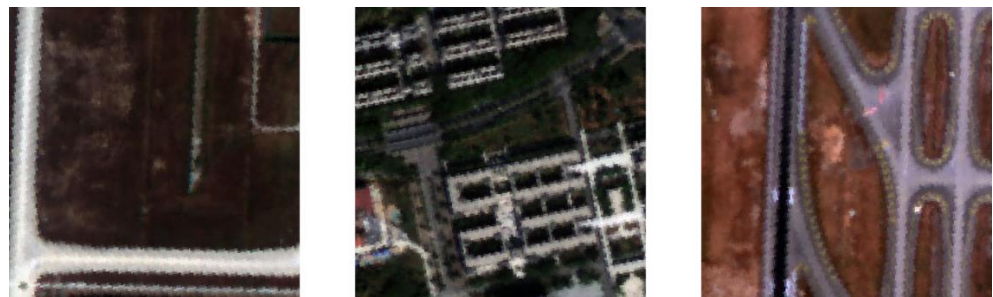


Figure 8. Sample images used in frequency-domain feature statistics.

It can be seen from Figure 9 that the distribution results of Bicubic, ESPCN, and ESRT are quite different from those of the HR images. The SD and IE distribution trends of SRCNN and SRGAN in different frequency bands were similar to those of the HR images, but the values were generally lower than those of the HR images. The SD and IE of the reconstructed feature map of the WGAN network were closer to the HR image feature map. It was proved that the WGAN can not only have a positive effect on the recovery of local high-frequency information, but also balance the information of different frequency bands to make it closer to the HR image.

To further verify the performance of the WGAN, Table 6 shows the different evaluation standards of the other five models. Due to the introduction of the WGAN, the evaluation index of the overall image was no longer only from the spatial gray distribution and information content, but also from the spectrum and overall reconstruction error. Therefore, the SAM and RMSE were introduced.

Table 6. Quality evaluation index results of different SR methods.

Method	Trained on Simulated Dataset				Trained on GF_Sen Dataset			
	PSNR	SSIM	SAM	RMSE	PSNR	SSIM	SAM	RMSE
Bicubic					18.863	0.593	0.301	30.039
SRCNN	17.975	0.506	0.284	28.506	18.191	0.512	0.316	30.553
ESPCN	18.006	0.612	0.302	29.864	18.266	0.582	0.303	29.864
ESRT	19.092	0.461	0.219	25.322	20.287	0.669	0.191	24.484
SRGAN	18.654	0.611	0.385	36.675	19.217	0.639	0.372	36.168
ESRGAN	18.975	0.621	0.286	28.948	19.458	0.647	0.338	33.060
WGAN	22.142	0.836	0.178	18.217	23.564	0.855	0.188	17.421

The analysis of the results indicated that the WGAN achieved a higher PSNR and SSIM performance on the GF_Sen. In addition, comparing the SAM and RMSE, we can find that simple or lightweight networks may reduce spectral distortion and minimize errors to a certain extent because they had less complexity in the operation of the image. However, the RMSE and SAM values of the reconstruction results of the WGAN showed the smallest values, which especially proved the effectiveness of using the wavelet transform into the network to utilize the frequency-domain features.

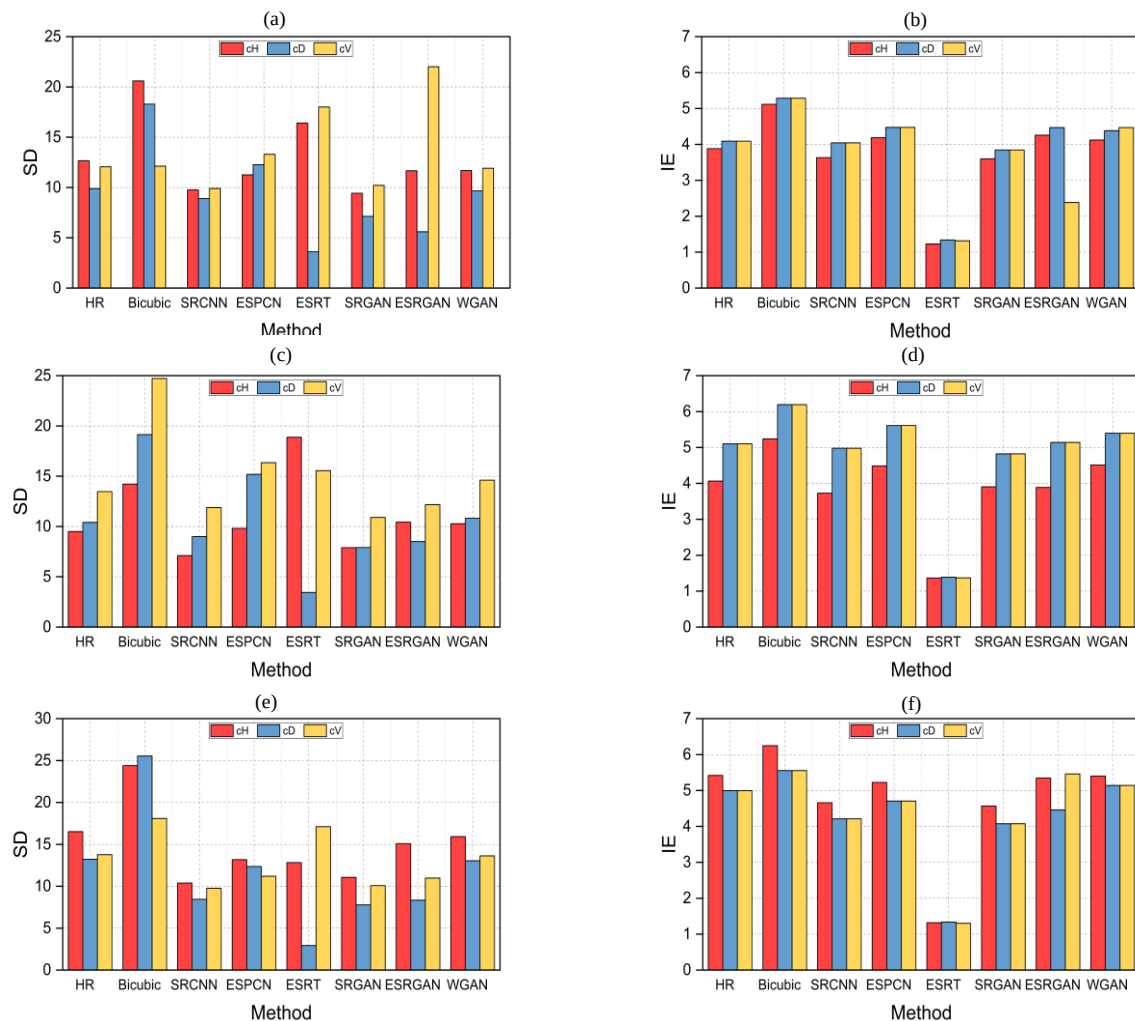


Figure 9. Evaluation histogram of high-frequency characteristic images with different reconstruction results. (a) and (b), (c) and (d), (e) and (f) are the results of pictures 1–3, respectively.

5. Discussion

5.1. Pros and Cons of CSWGAN

Currently, there is a need for satellite data with higher spatial and temporal resolution than what can be provided by individual imaging sensors, such as Sentinel-2 or GaoFen-2, in order to better understand surface changes and their causal factors. To address this, this study proposed a deep learning super-resolution reconstruction method called CSWGAN based on a cascaded structure, which downsampled Sentinel-2 images to a spatial resolution of 4 m, providing richer spatial details within a short time interval. The experimental results showed that, compared to representative works of different network structures, SRCNN and ESPCN, although having simple structures, showed a weaker performance for images with complex texture information. ESRT can achieve relatively good results in terms of metrics; however, it still tended to lose certain high-frequency information, and the perceptual quality of the reconstructed images still needs improvement. SRGAN

and ESRGAN performed relatively well in both perceptual and objective metrics. The CSWGAN, unlike other single-structure networks, leveraged a cascading structure that can harness the advantages of multiple individual networks, yielding superior perceptual performance and high-quality images. The proposed SGAN and WGAN have also been proven to show good results in the spatial and frequency domains, respectively. The experimental results demonstrated that the cascaded network structure combining spatial- and frequency-domain information can leverage the advantages of multiple single networks, resulting in superior perceptual performance and generating high-quality images. As the demand for precision increases, there are a growing number of networks in the field of computer vision with powerful image representation capabilities [44–46]. In the future, it is hoped that further research will be conducted on how to transfer these models to the application of remote sensing images. Loss function, as an important component guiding the network towards the correct direction, plays a crucial role in the allocation of coefficients for different loss functions to enhance the performance of the network. It is essential to scientifically and efficiently balance different loss terms to optimize the performance. Many studies on multi-task learning networks have proposed learning methods for loss function weights [47,48], and future attempts will also be made to refer to these algorithms to further improve the loss function.

In addition, the preliminary experimental results on network architecture still focus on displaying the RGB channels. However, this is limited for remote sensing images with multi-band characteristics [49]. For example, the vegetation red edge band of Sentinel-2 can effectively detect vegetation health [50], and the coastal/aerosol band (B1) can be used for monitoring water vapor and aerosol corrections [51,52]. Nevertheless, the spatial resolution of these bands cannot meet the actual application requirements. Therefore, the study of the super-reconstruction of multi-band cascaded spatial–frequency-domain network models will be part of future work. In contrast to Dong et al.’s [53], we find that conducting experiments on the RGB channels first is beneficial because most multi-band reconstruction methods are built upon the high-resolution RGB channels.

Finally, there are significant resolution differences among different sensors, such as a $3\times$ resolution difference between Landsat-7 and Sentinel-2, and a $2\times$ resolution difference between GaoFen-1 and GaoFen-2. In such cases, our model may require additional training to meet the super-resolution needs of different sensors. Furthermore, there can be large resolution differences between certain datasets, for example, at least an $8\times$ resolution difference between Landsat TM/ETM+ and MODIS, and a $30\times$ resolution difference between Sentinel-2 and Sentinel-3. In these cases, it is challenging to accurately handle the temporal, spatial, and spectral degradation relationships among multiple data using SR methods. Therefore, the use of spatiotemporal data fusion algorithms is vital to address the constraints of spatiotemporal resolution [54–57]. Based on our proposed model, a feasible improvement is to provide one or two pairs of high spatial resolution data sourced from low temporal resolution, and high temporal resolution data with low spatial resolution from prior dates as inputs to the model. This approach aims to predict one or more high temporal resolution data with low spatial resolution on prediction dates. The effectiveness of this plan will be discussed in the future.

5.2. Application of GF_Sen

In this paper, we conducted a study on the GF_Sen dataset, simulated dataset, and multi-scene dataset WHU, among others, and compared the accuracy of the four most mature models, as well as two single models internally, to gain an insight into their performance. In recent years, there have been many advanced deep learning-based super-resolution reconstruction methods. In the future, there is an interest in conducting more comparisons to explore the advantages and disadvantages of different datasets for different models. Although the CSWGAN is trained based on cross-sensor datasets, for the simulated dataset, this paper reconstructed scenes from the WHU dataset and achieved better results than other models for most scenes, demonstrating the universality of the

model. The proposal of the cross-sensor remote sensing dataset GF_Sen based on GaoFen-2 and Sentinel-2 supplements the current benchmark pairs of remote sensing image datasets in the field of remote sensing super-resolution reconstruction, providing a benchmark dataset for applications such as fine-grained classification [58], building extraction [59], tree detection [60], and detailed land cover mapping [61]. This dataset addresses the conditions of (1) different sensors, (2) regional diversity, (3) temporal consistency, and (4) challenging scenes, supporting the design of more advanced SR models.

6. Conclusions

In this paper, we introduced a new real high- and low-resolution dataset (GF_Sen) and proposed a cascade structure network CSWGAN for super-resolution.

GF_Sen is a benchmark paired dataset of real remote sensing images, which is composed of Sentinel-2 and GaoFen-2 images. The experiments demonstrate that networks trained on this dataset can have a better performance than those trained on a simulated dataset.

In addition, a GAN-based remote sensing image super-resolution method CSWGAN is proposed in this paper, which adopts a cascaded generative adversarial network and is composed of single-structure networks proposed by the SGAN and WGAN. The CSWGAN combines the advantages of the SGAN and WGAN and has strengthened its capability to model long-range dependencies between different frequency domains and regions, as well as acquiring global topological information and local texture details. This solves the challenge of using a single network model to simultaneously target real low-resolution images and obtain better reconstruction results. Finally, comparing with different networks and different scene types and scales of scenes, it has been proven that the CSWGAN has a more obvious advantage in reconstructing scenes with more detailed textures, and achieves the best quantitative and qualitative results on relevant datasets.

Author Contributions: X.W., Z.A. and Y.F. conceived and designed the study and methods; X.W. and R.L. analyzed the data; X.W., Y.X. and Y.G. wrote the paper, and all co-authors contributed to the interpretation of the results and to the text. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China, grant number 42071399, Ministry of education of Humanities and Social Science project under Grant [number 23YJAZH019], and Tibet Autonomous Region Science and Technology Program under Grant [number XZ202301ZY0021G].

Data Availability Statement: The manuscript dataset is currently undergoing a patent application process, and after the patent application is completed, we will share the data and code.

Acknowledgments: We are grateful to the editor and anonymous reviewers for their valuable comments on this manuscript.

Conflicts of Interest: Author Runhao Li was employed by the company China Water Resources Pearl River Planning, Surveying & Designing Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Ienco, D.; Interdonato, R.; Gaetano, R.; Ho Tong Minh, D. Combining Sentinel-1 and Sentinel-2 Satellite Image Time Series for land cover mapping via a multi-source deep learning architecture. *ISPRS J. Photogramm. Remote Sens.* **2019**, *158*, 11–22. [\[CrossRef\]](#)
2. Tang, X.; Bratley, K.H.; Cho, K.; Bullock, E.L.; Olofsson, P.; Woodcock, C.E. Near real-time monitoring of tropical forest disturbance by fusion of Landsat, Sentinel-2, and Sentinel-1 data. *Remote Sens. Environ.* **2023**, *294*, 113626. [\[CrossRef\]](#)
3. Pelletier, F.; Cardille, J.A.; Wulder, M.A.; White, J.C.; Hermosilla, T. Inter- and intra-year forest change detection and monitoring of aboveground biomass dynamics using Sentinel-2 and Landsat. *Remote Sens. Environ.* **2024**, *301*, 113931. [\[CrossRef\]](#)
4. Hafner, S.; Ban, Y.; Nascetti, A. Unsupervised domain adaptation for global urban extraction using Sentinel-1 SAR and Sentinel-2 MSI data. *Remote Sens. Environ.* **2022**, *280*, 113192. [\[CrossRef\]](#)

5. Zhou, H.; Liu, S.; Mo, X.; Hu, S.; Zhang, L.; Ma, J.; Bandini, F.; Grosen, H.; Bauer-Gottwein, P. Calibrating a hydrodynamic model using water surface elevation determined from ICESat-2 derived cross-section and Sentinel-2 retrieved sub-pixel river width. *Remote Sens. Environ.* **2023**, *298*, 113796. [\[CrossRef\]](#)
6. Ren, B.; Ma, S.; Hou, B.; Hong, D.; Chanussot, J.; Wang, J.; Jiao, L. A dual-stream high resolution network: Deep fusion of GF-2 and GF-3 data for land cover classification. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102896. [\[CrossRef\]](#)
7. Wu, Q.; Zhong, R.; Zhao, W.; Song, K.; Du, L. Land-cover classification using GF-2 images and airborne lidar data based on Random Forest. *Int. J. Remote Sens.* **2018**, *40*, 2410–2426. [\[CrossRef\]](#)
8. Farsiu, S.; Robinson, D.; Elad, M.; Milanfar, P. Advances and challenges in super-resolution. *Int. J. Imaging Syst. Technol.* **2004**, *14*, 47–57. [\[CrossRef\]](#)
9. 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. [\[CrossRef\]](#)
10. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Trans. Acoust. Speech Signal Process.* **1981**, *29*, 1153–1160. [\[CrossRef\]](#)
11. Kim, J.; Lee, J.K.; Lee, K.M. Accurate Image Super-Resolution Using Very Deep Convolutional Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1646–1654.
12. Pashaie, M.; Starek, M.J.; Kamangir, H.; Berryhill, J. Deep Learning-Based Single Image Super-Resolution: An Investigation for Dense Scene Reconstruction with UAS Photogrammetry. *Remote Sens.* **2020**, *12*, 1757. [\[CrossRef\]](#)
13. Huan, H.; Li, P.; Zou, N.; Wang, C.; Xie, Y.; Xie, Y.; Xu, D. End-to-End Super-Resolution for Remote-Sensing Images Using an Improved Multi-Scale Residual Network. *Remote Sens.* **2021**, *13*, 666. [\[CrossRef\]](#)
14. Dong, R.; Mou, L.; Zhang, L.; Fu, H.; Zhu, X.X. Real-world remote sensing image super-resolution via a practical degradation model and a kernel-aware network. *ISPRS J. Photogramm. Remote Sens.* **2022**, *191*, 155–170. [\[CrossRef\]](#)
15. Panagiotopoulou, A.; Grammatikopoulos, L.; Charou, E.; Bratsolis, E.; Petrogonas, J. Very Deep Super-Resolution of Remotely Sensed Images with Mean Square Error and Var-norm Estimators as Loss Functions. *arXiv* **2020**, arXiv:2007.15417.
16. Wang, X.; Wu, Y.; Ming, Y.; Lv, H. Remote Sensing Imagery Super Resolution Based on Adaptive Multi-Scale Feature Fusion Network. *Sensors* **2020**, *20*, 1142. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wang, Y.; Shao, Z.; Lu, T.; Liu, L.; Huang, X.; Wang, J.; Jiang, K.; Zeng, K. A lightweight distillation CNN-transformer architecture for remote sensing image super-resolution. *Int. J. Digit. Earth* **2023**, *16*, 3560–3579. [\[CrossRef\]](#)
18. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
20. Lu, T.; Wang, J.; Zhang, Y.; Wang, Z.; Jiang, J. Satellite Image Super-Resolution via Multi-Scale Residual Deep Neural Network. *Remote Sens.* **2019**, *11*, 1588. [\[CrossRef\]](#)
21. 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. [\[CrossRef\]](#)
22. Ao, Z.; Wu, F.; Hu, S.; Sun, Y.; Su, Y.; Guo, Q.; Xin, Q. Automatic segmentation of stem and leaf components and individual maize plants in field terrestrial LiDAR data using convolutional neural networks. *Crop J.* **2022**, *10*, 1239–1250. [\[CrossRef\]](#)
23. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In Proceedings of the Computer Vision—ECCV 2016, Cham, Switzerland, 8–16 October 2016; pp. 694–711.
24. Mathieu, M.; Couprie, C.; Lecun, Y. Deep multi-scale video prediction beyond mean square error. *arXiv* **2016**, arXiv:1511.05440.
25. Guo, D.; Xia, Y.; Xu, L.; Li, W.; Luo, X. Remote sensing image super-resolution using cascade generative adversarial nets. *Neurocomputing* **2021**, *443*, 117–130. [\[CrossRef\]](#)
26. Ledig, C.; Theis, L.; Huszár, 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 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 105–114.
27. 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 Computer Vision—ECCV 2018 Workshops, Cham, Switzerland, 8–14 September 2019; pp. 63–79.
28. Peng, D.; Yang, W.; Liu, C.; Lü, S. SAM-GAN: Self-Attention supporting Multi-stage Generative Adversarial Networks for text-to-image synthesis. *Neural Netw.* **2021**, *138*, 57–67. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Zong, L.; Chen, L. Single Image Super-Resolution Based on Self-Attention. In Proceedings of the 2019 IEEE International Conference on Unmanned Systems and Artificial Intelligence (ICUSAI), Xi'an, China, 22–24 November 2019; pp. 56–60.
30. Lu, Z.; Liu, H.; Li, J.; Zhang, L. Efficient Transformer for Single Image Super-Resolution. *arXiv* **2021**, arXiv:2108.11084.
31. Li, J.; Meng, Y.; Tao, C.; Zhang, Z.; Yang, X.; Wang, Z.; Wang, X.; Li, L.; Zhang, W. ConvFormerSR: Fusing Transformers and Convolutional Neural Networks for Cross-Sensor Remote Sensing Imagery Super-Resolution. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–15. [\[CrossRef\]](#)
32. Chan, R.; Chan, T.; Shen, L.; Shen, Z. Wavelet Algorithms for High-Resolution Image Reconstruction. *SIAM J. Sci. Comput.* **2004**, *24*, 1408–1432. [\[CrossRef\]](#)

33. Kinebuchi, K.; Muresan, D.D.; Parks, T.W. Image interpolation using wavelet based hidden Markov trees. In Proceedings of the Acoustics, Speech, and Signal Processing, 2001 on IEEE International Conference, Salt Lake City, UT, USA, 7–11 May 2001; Volume 3, pp. 1957–1960.
34. Zhou, R.; Lahoud, F.; Helou, M.; Süssstrunk, S. A comparative study on wavelets and residuals in deep super resolution. *Electron. Imaging* **2019**, *2019*, 135–1–135–7. [\[CrossRef\]](#)
35. Guo, T.; Mousavi, H.S.; Vu, T.H.; Monga, V. Deep Wavelet Prediction for Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1100–1109.
36. Xue, S.; Qiu, W.; Liu, F.; Jin, X. Wavelet-based residual attention network for image super-resolution. *Neurocomputing* **2020**, *382*, 116–126. [\[CrossRef\]](#)
37. Feng, X.; Zhang, W.; Su, X.; Xu, Z. Optical Remote Sensing Image Denoising and Super-Resolution Reconstructing Using Optimized Generative Network in Wavelet Transform Domain. *Remote Sens.* **2021**, *13*, 1858. [\[CrossRef\]](#)
38. Huang, H.; He, R.; Sun, Z.; Tan, T. Wavelet-SRNet: A Wavelet-Based CNN for Multi-scale Face Super Resolution. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1698–1706.
39. 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.
40. Sheng, G.; Yang, W.; Xu, T.; Sun, H. High-resolution satellite scene classification using a sparse coding based multiple feature combination. *Int. J. Remote Sens.* **2012**, *33*, 2395–2412. [\[CrossRef\]](#)
41. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced Deep Residual Networks for Single Image Super-Resolution. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 1132–1140.
42. Fernandez-Beltran, R.; Carmona, P.; Pla, F. Single-frame super-resolution in remote sensing: A practical overview. *Int. J. Remote Sens.* **2017**, *38*, 314–354. [\[CrossRef\]](#)
43. Karathanassi, V.; Kolokousis, P.; Ioannidou, S. A comparison study on fusion methods using evaluation indicators. *Int. J. Remote Sens.* **2007**, *28*, 2309–2341. [\[CrossRef\]](#)
44. Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; Yu, F. Dual Aggregation Transformer for Image Super-Resolution. *arXiv* **2023**, arXiv:2308.03364. [\[CrossRef\]](#)
45. Li, B.; Li, X.; Zhu, H.; Jin, Y.; Feng, R.; Zhang, Z.; Chen, Z. SeD: Semantic-Aware Discriminator for Image Super-Resolution. *arXiv* **2024**, arXiv:2402.19387.
46. Gandikota, K.V.; Chandramouli, P. Text-guided Explorable Image Super-resolution. *arXiv* **2024**, arXiv:2403.01124.
47. Kendall, A.; Gal, Y.; Cipolla, R. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
48. Yang, B.; Xiang, X.; Kong, W.; Peng, Y.; Yao, J. Adaptive multi-task learning using lagrange multiplier for automatic art analysis. *Multimed. Tools Appl.* **2022**, *81*, 3715–3733. [\[CrossRef\]](#)
49. Sung Cheol, P.; Min Kyu, P.; Moon Gi, K. Super-resolution image reconstruction: A technical overview. *IEEE Signal Process. Mag.* **2003**, *20*, 21–36. [\[CrossRef\]](#)
50. Shiklomanov, A.N.; Dietze, M.C.; Viskari, T.; Townsend, P.A.; Serbin, S.P. Quantifying the influences of spectral resolution on uncertainty in leaf trait estimates through a Bayesian approach to RTM inversion. *Remote Sens. Environ.* **2016**, *183*, 226–238. [\[CrossRef\]](#)
51. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [\[CrossRef\]](#)
52. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. [\[CrossRef\]](#)
53. Dong, X.; Wang, L.; Sun, X.; Jia, X.; Gao, L.; Zhang, B. Remote Sensing Image Super-Resolution Using Second-Order Multi-Scale Networks. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 3473–3485. [\[CrossRef\]](#)
54. Zhu, X.; Helmer, E.; Gao, F.; Liu, D.; Chen, J.; Lefsky, M. A flexible spatiotemporal method for fusing satellite images with different resolutions. *Remote Sens. Environ.* **2016**, *172*, 165–177. [\[CrossRef\]](#)
55. Zhang, H.; Sun, Y.; Shi, W.; Guo, D.; Zheng, N. An object-based spatiotemporal fusion model for remote sensing images. *Eur. J. Remote Sens.* **2021**, *54*, 86–101. [\[CrossRef\]](#)
56. Li, J.; Li, Y.; He, L.; Chen, J.; Plaza, A. Spatio-temporal fusion for remote sensing data: An overview and new benchmark. *Sci. China Inf. Sci.* **2020**, *63*, 140301. [\[CrossRef\]](#)
57. Wang, Q.; Atkinson, P.M. Spatio-temporal fusion for daily Sentinel-2 images. *Remote Sens. Environ.* **2018**, *204*, 31–42. [\[CrossRef\]](#)
58. Fu, G.; Liu, C.; Zhou, R.; Sun, T.; Zhang, Q. Classification for High Resolution Remote Sensing Imagery Using a Fully Convolutional Network. *Remote Sens.* **2017**, *9*, 498. [\[CrossRef\]](#)
59. Li, W.; He, C.; Fang, J.; Zheng, J.; Fu, H.; Yu, L. Semantic Segmentation-Based Building Footprint Extraction Using Very High-Resolution Satellite Images and Multi-Source GIS Data. *Remote Sens.* **2019**, *11*, 403. [\[CrossRef\]](#)

60. Zheng, J.; Fu, H.; Li, W.; Wu, W.; Yu, L.; Yuan, S.; Tao, W.Y.W.; Pang, T.K.; Kanniah, K.D. Growing status observation for oil palm trees using Unmanned Aerial Vehicle (UAV) images. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 95–121. [[CrossRef](#)]
61. Dong, R.; Li, C.; Fu, H.; Wang, J.; Li, W.; Yao, Y.; Gan, L.; Yu, L.; Gong, P. Improving 3-m Resolution Land Cover Mapping through Efficient Learning from an Imperfect 10-m Resolution Map. *Remote Sens.* **2020**, *12*, 1418. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.