remote sensing

MDPI

Article

Single-Image Super-Resolution Method for Rotating Synthetic Aperture System Using Masking Mechanism

Yu Sun [1], Xiyang Zhi [1,*], Shikai Jiang [1], Tianjun Shi [1], Jiachun Song [1], Jiawei Yang [1], Shengao Wang [2] and Wei Zhang [1]

[1] Research Center for Space Optical Engineering, Harbin Institute of Technology, Harbin 150001, China; ysun@stu.hit.edu.cn (Y.S.); jiangshikai@hit.edu.cn (S.J.); shitianjun@stu.hit.edu.cn (T.S.); 22s021001@stu.hit.edu.cn (J.S.); 23s021001@stu.hit.edu.cn (J.Y.)
[2] Division of System Engineering, Boston University, Boston, MA 02215, USA; wsashawn@bu.edu
* Correspondence: zhixiyang@hit.edu.cn; Tel.: +86-0451-86414883

Abstract: The emerging technology of rotating synthetic aperture (RSA) presents a promising solution for the development of lightweight, large-aperture, and high-resolution optical remote sensing systems in geostationary orbit. However, the rectangular shape of the primary mirror and the distinctive imaging mechanism involving the continuous rotation of the mirror lead to a pronounced decline in image resolution along the shorter side of the rectangle compared to the longer side. The resolution also exhibits periodic time-varying characteristics. To address these limitations and enhance image quality, we begin by analyzing the imaging mechanism of the RSA system. Subsequently, we propose a single-image super-resolution method that utilizes a rotated varied-size window attention mechanism instead of full attention, based on the Vision Transformer architecture. We employ a two-stage training methodology for the network, where we pre-train it on images masked with stripe-shaped masks along the shorter side of the rectangular pupil. Following that, we fine-tune the network using unmasked images. Through the strip-wise mask sampling strategy, this two-stage training approach effectively circumvents the interference of lower confidence (clarity) information and outperforms training the network from scratch using the unmasked degraded images. Our digital simulation and semi-physical imaging experiments demonstrate that the proposed method achieves satisfactory performance. This work establishes a valuable reference for future space applications of the RSA system.

Keywords: optical remote sensing; super-resolution (SR); rotating synthetic aperture; masked autoencoder; vision transformer; rectangular pupil

Citation: Sun, Y.; Zhi, X.; Jiang, S.; Shi, T.; Song, J.; Yang, J.; Wang, S.; Zhang, W. Single-Image Super-Resolution Method for Rotating Synthetic Aperture System Using Masking Mechanism. *Remote Sens.* **2024**, *16*, 1508. https://doi.org/10.3390/rs16091508

Academic Editor: Dusan Gleich

Received: 13 March 2024
Revised: 22 April 2024
Accepted: 23 April 2024
Published: 25 April 2024

1. Introduction

Optical remote sensing satellites in geostationary orbit offer the ability for high spatial and temporal resolution, making them a crucial part of space-based observation technology [1–6]. Compared to low-Earth-orbit optical remote sensing satellites, they provide several advantages: (1) they remain stationary relative to the observed area, allowing for observations of target regions over extended periods with higher temporal resolution; (2) by adjusting their direction, they can quickly acquire observational images for the corresponding area, making them particularly suitable for emergency response tasks; (3) flexible satellite mission planning enables repeated observations of multiple hotspots and large areas; and (4) compared to line-scan satellites, the array camera has the advantage of a long integration time, resulting in the acquisition of high-quality images.

Due to the high orbit altitude and the fact that the imaging object distance is tens of times greater than that of a low Earth orbit, high-orbit remote sensing satellites require a larger aperture to ensure imaging quality. Various techniques have been developed to overcome the aperture limitations, such as segmented mirror technology, membrane diffraction

imaging technology, optical synthetic aperture technology, and rotating synthetic aperture (RSA) technology. The segmented mirror technology employs segmented sub-mirrors to assemble into a large aperture primary mirror. The sub-mirrors are folded during launch and expanded after entering orbit. During the imaging process, it is imperative to guarantee that each sub-mirror is accurately assembled. The primary mirror expansion and support structure's complexity result in elevated expenses [7,8]. The membrane diffraction imaging technology utilizes thin-film materials to create imaging devices. Under the precondition of achieving the same resolution ability, the system mass is only one-seventh of that of a traditional large-aperture single-reflection mirror system, considerably decreasing the rocket-carrying capacity requirements. The membrane mirror's surface necessitates lower precision than a traditional reflection mirror, which reduces manufacturing difficulty and enables mass production, with potential for significantly lower costs. However, this system has the disadvantage of color dispersion, with a narrow spectral response range of only approximately 40 nm, and lower diffraction efficiency, which limits the system's practical application to some extent [9–11]. The optical synthetic aperture imaging technology is based on interferometric imaging principles, employing small-aperture systems to synthesize a large-aperture system. Similar to the segmented mirror technology, its advantage is avoiding the processing of large-aperture lenses and reducing launch costs by using small-aperture systems. However, the synthesis of the large-aperture must come at the expense of reducing light flux, resulting in a reduced signal-to-noise ratio. This system must satisfy the co-phasing condition to attain ideal imaging, and therefore error monitoring and precise phase adjustment make the engineering implementation of the system extremely challenging [12–14]. Compared to the methods mentioned above, the RSA system presents a more advanced alternative. It employs a primary mirror that is rotatable and possesses a large aspect ratio, as illustrated in Figure 1. Through the rotation of the primary mirror, a sequence of images is generated, capturing high-resolution information from various directions. However, this also results in significantly higher resolution along the long edge compared to the short edge [15–18]. Furthermore, the rotation generates periodic fluctuations in image quality, making it essential to employ image enhancement techniques to improve the quality of the imaging system.

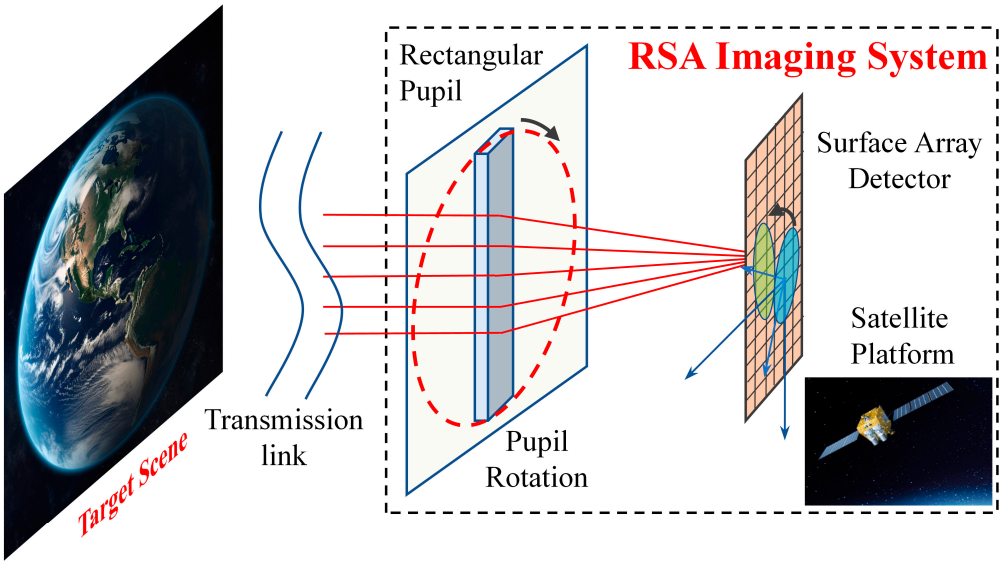


Figure 1. Imaging principle of the rotating synthetic aperture (RSA) system. Blue arrows represent the coordinate axes of the detector plane, while black arrows indicate the rotation of the pupil.

The task of single-image super-resolution (SISR) is to reconstruct a high-resolution image (HR) from a corresponding low-resolution image (LR). The relationship between these two quantities is described by the classical degradation model, $\mathrm{LR} = (\mathrm{HR} * k)_{\downarrow} + n$,

where k represents a blur kernel, and $*$, n, and $\downarrow$ denote the convolution operator, additive noise, and down-sampling operator, respectively. Deep learning-based SISR methods have demonstrated remarkable performance compared to traditional methods, and have become the prevailing approach in recent years [19–22]. These methods can be broadly classified into two categories: explicit methods based on the classical degradation model or its variants, and implicit methods that leverage the data distribution within the training dataset [23–25]. Explicit methods aim to directly learn the blur kernel k and additive noise n in the classical degradation model from the training data. Representative approaches include SRGAN [26], EDSR [27], SRMD [28], and Real-ESRGAN [29]. Another group of explicit methods, such as KernelGAN [30], DualSR [31], and DBPI [32], rely on the internal statistics of patch recurrence. However, the RSA system's point spread function (PSF) continuously changes during the imaging process due to the rotation of the primary mirror, resulting in images with temporal periodicity and spatial asymmetry. Existing explicit methods do not account for this special characteristic of the blur kernel, leading to suboptimal performance. On the other hand, implicit methods do not rely on any explicit parameterization and instead typically learn the underlying super-resolution (SR) model implicitly through the data distribution within training datasets. Representative approaches include CinCGAN [33] and FSSR [34]. The general meanings and types of the methods mentioned above are depicted in Table 1. However, the RSA system presents a challenge for implicit methods due to the existence of multiple degraded images of the same target scene resulting from different rotation angles, which makes the data distribution more complex. Furthermore, the methods mentioned above mostly rely on convolutional neural networks (CNNs), but the strong long-range dependency of remote sensing images makes it difficult for CNNs with local inductive bias to meet application requirements. In summary, the RSA system possesses unique imaging characteristics that make it difficult to apply existing SISR methods directly. Therefore, it is crucial to conduct research on targeted remote sensing image SR methods based on the degradation characteristics of the system.

Table 1. Overview of deep learning-based single-image super-resolution methods.

Method	General Meaning	Type
SRGAN [26]	Generative adversarial networks for image super-resolution	Explicit methods rely on external training datasets
EDSR [27]	Enhanced deep residual networks for image super-resolution	
SRMD [28]	Image super-resolution networks for multiple degradations	
Real-ESRGAN [29]	Real-world enhanced super-resolution generative adversarial networks	
KernelGAN [30]	Generative adversarial networks for kernel estimation	Explicit methods rely on internal statistics
DualSR [31]	Dual learning for image super-resolution	
DBPI [32]	Dual back-projection-based internal learning	
CinCGAN [33]	Cycle-in-cycle generative adversarial networks for image super-resolution	Implicit methods
FSSR [34]	Frequency separation for image super-resolution	

To address this challenge, we begin by examining the non-circular symmetry spatial distribution and temporal variability of the PSF in relation to the imaging mechanism of the RSA system. Subsequently, we propose an SISR method based on Vision Transformer (ViT) [35], which is trained in a two-stage process. In the first stage, we pre-train the network on degraded images masked along the short edge direction of the rectangular primary mirror. Then, in the second stage, we perform fine-tuning using unmasked images. This approach proves to be superior to directly training the network on the original degraded images without masking. Given the reduced resolution along the shorter edge direction caused by the non-circular symmetry of the pupil, our primary objective is to compensate for the substantial loss of information along the shorter side direction, thereby enhancing the resolution in that direction. Building upon this, the proposed method can further achieve an increase in resolution across all directions. Additionally, considering the specific characteristics of remote sensing images, we replace the original Vision Transformer blocks with rotated varied-size window-based attention blocks [36]. These blocks introduce local windows with different locations, sizes, shapes, and angles to calculate window-based

attention. Finally, we validate the effectiveness of our proposed method through digital simulations experiments as well as semi-physical imaging experiments.

2. Materials and Methods

2.1. Analysis of Imaging Mechanism of the RSA System

The RSA system can be regarded as a diffraction-limited system, in which any point source (x_0, y_0) on the object plane emits a diverging spherical wave that projects onto the entrance pupil. This wave is then transformed by the system into a converging spherical wave on the exit pupil, which projects onto the image plane at position (x_i, y_i). In the case of polychromatic illumination, the diffraction-limited incoherent imaging system is a linear space-invariant system of intensity transformation. The object–image relationship can be expressed as (ignoring the constant coefficient):

$$I_i(x_i, y_i) = \iint_{-\infty}^{\infty} I_g(x_0, y_0) h_I(x_i - x_0, y_i - y_0) dx_0 dy_0 \tag{1}$$

where I_i represents the intensity distribution of the image plane, I_g represents the intensity distribution of the geometrical optics ideal image, and h_I represents the intensity impulse response, i.e., the PSF, which denotes the intensity distribution of diffraction spots produced by point objects.

The equation above demonstrates that when a point source is used as the input elemental object, it generates an image spot on the image plane with the ideal image point of geometrical optics at its center. The intensity distribution of the image plane results from the superposition of the image spots produced by all point sources on the object plane. The shape of the image spot is described by the PSF. According to Fresnel–Kirchhoff's diffraction formula, the PSF is obtained by subjecting the pupil function to Fourier transform and squaring the resulting modulus. For the RSA system, more specifically, its pupil function at time t can be expressed as:

$$P(x, y, t) = \text{rect}\left(\frac{x\cos(wt + \varphi_0) - y\sin(wt + \varphi_0)}{a}\right)\text{rect}\left(\frac{x\sin(wt + \varphi_0) + y\cos(wt + \varphi_0)}{b}\right) \tag{2}$$

where a and b are the length and width of the rectangle, respectively, w is the angular velocity of the rectangular primary mirror rotation, φ_0 is the initial phase, and $\text{rect}(\cdot)$ represents the rectangle function, which is defined as:

$$\text{rect}(s) = \begin{cases} 1 \text{ if} |s| < 0.5 \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

The PSF can be obtained according to the pupil function as follows:

$$\begin{aligned} \text{PSF}(x, y, t) = \quad & ab\,\text{sin c}(a(x\cos(wt + \varphi_0) - y\sin(wt + \varphi_0))) \\ & \times \text{sin c}(b(x\sin(wt + \varphi_0) + y\cos(wt + \varphi_0))) \end{aligned} \tag{4}$$

Figure 2 reveals that the PSF takes on an elliptical shape if the secondary diffraction effect is not considered. The length-to-width ratio of the primary mirror's rectangular shape determines the form of this ellipse, with the orientation of the longer axis aligned with the shorter side of the rectangle [15].

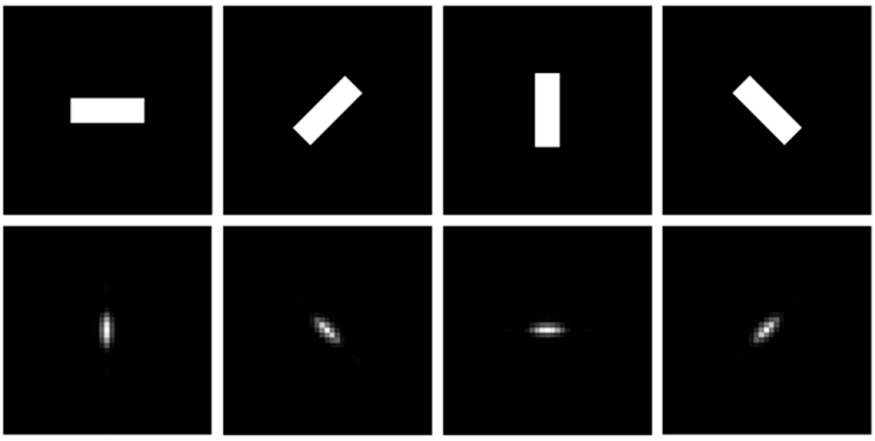


Figure 2. Rectangular pupils with different rotation angles and their point spread functions: 0°, 45°, 90°, and 135°, from left to right.

2.2. Overview of the Image Super-Resolution Approach

As stated in Section 1, the rotation of the rectangular pupil introduces temporal variation in the blur kernel, posing a challenge for accurately estimating the blur kernel by the model. Additionally, the many-to-one relationship between LR and HR images increases the difficulty of learning the data distribution for implicit methods. In other words, the unique imaging mechanism of the RSA system presents obstacles for the deep neural network to acquire the desired ability, which involves leveraging the high-resolution information preserved in the image itself in specific directions for super-resolution reconstruction in lower-resolution directions to significantly enhance the image resolution in the short side direction of the rectangle. Consequently, these challenges often result in the phenomenon of uneven resolution in the output reconstructed image, which is not significantly improved. Nevertheless, the time-sequential imaging method employed in the RSA system offers a benefit: the images captured within one rotation cycle of the rectangular pupil contain high-resolution information from various directions. Motivated by the work of [37], our training methodology involves initially masking the pixels along the shorter side direction of the rectangle and using a ViT-based super-resolution network to reconstruct them. This approach aims to effectively leverage the complementary information available across images captured at various rotation angles of the pupil. Subsequently, we fine-tune the model using the unmasked images. For the SR reconstruction module implementation, we employ the sub-pixel convolution layer [38] to upsample the features outputted by the decoder. The overall process is depicted in Figure 3. We have observed that this two-stage training approach significantly improves performance. Based on this observation, we believe that masking the pixels along the lower resolution direction can guide the network to focus more on the high-resolution information preserved along the longer edge of the rectangle. Furthermore, considering the presence of objects with varying orientations and scales in the remote sensing images obtained by the RSA system, we replace global self-attention with rotated varied-size window-based attention. This modified attention mechanism introduces shift, scale, and rotation parameters to the original window-based attention, enabling diversified windows of different locations, sizes, shapes, and angles to better handle objects with varying orientations and scales.

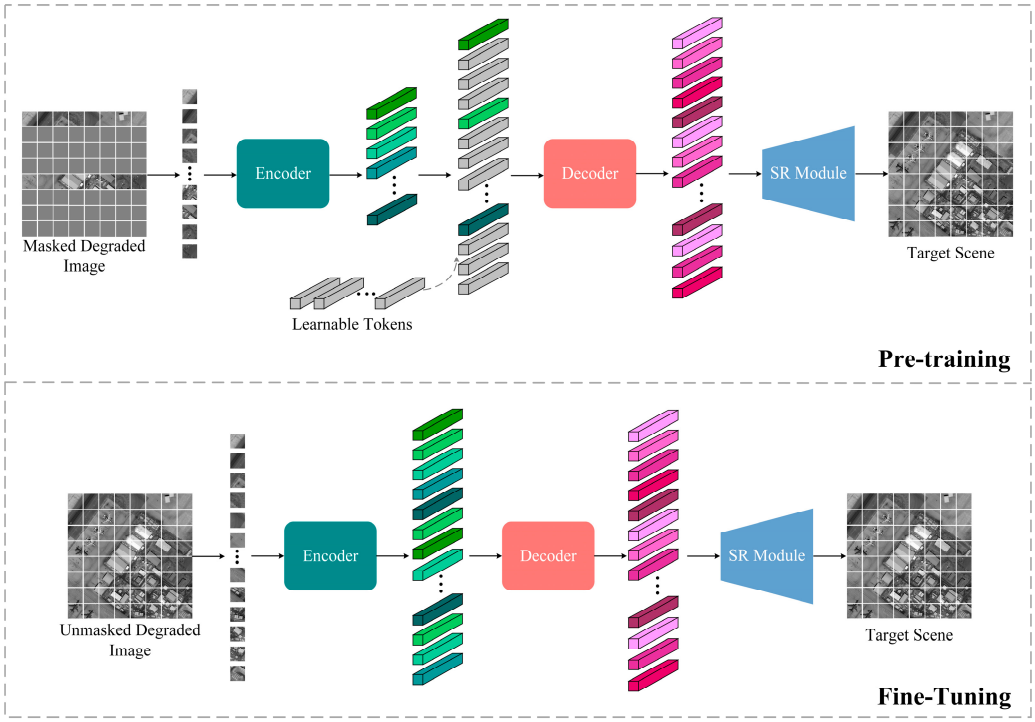


Figure 3. Overall framework of pre-training and fine-tuning. Different colors are used to distinguish between various modules and their outputs, while varying shades are employed to delineate the features extracted from different patches.

2.3. Encoder

The encoder in our model comprises a stack of rotated varied-size window-based attention blocks, as illustrated in Figure 4. The feed-forward network is a 2-layer multilayer perceptron with nonlinear activation functions in between. And the rotated varied-size multi-head attention will be detailed in Section 2.3.2.

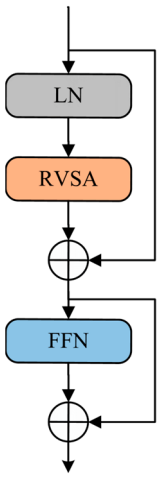


Figure 4. The structures of the rotated varied-size window-based attention block. It includes layer normalization (LN) [39], rotated varied-size multi-head attention (RVSA), and a feed-forward network (FFN).

2.3.1. Masked Autoencoder

The masked autoencoder (MAE) was proposed by He et al. with the aim of recovering masked parts of an image in the pixel space through an encoder–decoder structure, given the visible parts of the image [37]. This process involves partitioning the input image

into non-overlapping patches, followed by masking some patches according to a predetermined ratio. These masked patches are then treated as regions to be reconstructed. The original MAE utilizes random sampling to select masked patches, whereas we employ a corresponding mask sampling strategy that accounts for the unique pupil shape of the RSA system. Our mask sampling strategy involves masking pixels in a striped pattern along the direction of lower resolution, which is the shorter side of the rectangular pupil. This strategy serves to remove some of the lower confidence (clarity) priors by masking more low-resolution information. It can guide the model to focus more on and utilize high-resolution information along other directions, thereby avoiding interference caused by the specific degradation process resulting from the asymmetric characteristic of the PSF. As a simple example, let us consider the two perpendicular directions of edges shown in Figure 5. By masking along the lower-resolution direction, blurred edges are masked while preserving more pixels on the sharper edges along the longer side of the rectangle. This can mitigate the interference of partially blurred pixels, helping the model to reconstruct sharper super-resolution results and reduce the uneven resolution phenomenon more effectively. Additionally, ViT has a class token for classification, but since our task is image enhancement rather than classification, we do not use this token. This is another difference between our method and the original MAE. According to [37], even without the class token (with average pooling), the encoder can still work well.

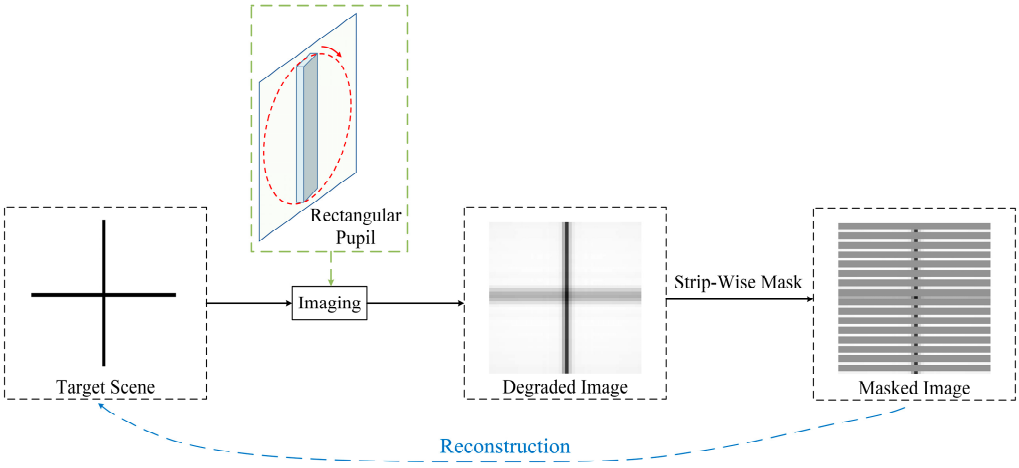


Figure 5. Explanation of mask sampling strategy.

2.3.2. Rotated Varied-Size Window-Based Attention

Unlike convolution, which possesses a local inductive bias, Transformer can adopt a more global perspective and exploit correlations between pixels of images through the attention mechanism. This not only inherently resonates with the goal of leveraging high-resolution information in specific directions to enhance resolution along the shorter side of the rectangle but also facilitates the utilization of long-range dependencies in images captured by the RSA system. In order to reduce computational complexity, ViT mostly uses window-based attention that differs from standard global self-attention by employing both a local attention mechanism and a window transfer mechanism [40]. In window-based attention, the fundamental processing unit is the patch. The network first conducts a patch partition operation on the input image, which involves partitioning the input image into non-overlapping patches. Specifically, given an input $X \in \mathbb{R}^{H\times W\times C}$ of size $H \times W \times C$ (where H, W, and C represent the width, height, and number of channels of the feature map, respectively), window-based attention first reshapes the input by partitioning it into $M\times M$ non-overlapping local windows, denoted as $X \in \mathbb{R}^{\frac{H}{M}\times\frac{W}{M}\times M^2\times C}$, where $\frac{HW}{M^2}$ is the total number of windows. For each window, the input features are denoted as $X_w \in \mathbb{R}^{M^2\times C}$, and thus, all input features can be represented as $\left\{X_{w_i}|i=1,\cdots,\frac{HW}{M^2}\right\}$. Following this,

standard multi-head self-attention is computed for each window. Let h denote the number of heads; the *query*, *key* and *value* matrices are represented by Q_w, K_w and V_w, respectively:

$$\left\{ \mathrm{Q}_{w_i}^{(j)} | i = 1, \cdots, \frac{HW}{M^2}, j = 1, \cdots, h \right\} \tag{5}$$

$$\left\{ \mathrm{K}_{w_i}^{(j)} | i = 1, \cdots, \frac{HW}{M^2}, j = 1, \cdots, h \right\} \tag{6}$$

$$\left\{ \mathrm{V}_{w_i}^{(j)} | i = 1, \cdots, \frac{HW}{M^2}, j = 1, \cdots, h \right\} \tag{7}$$

where i indexes the window and j indexes the head.

The attention calculations are performed within each non-overlapping local window:

$$\mathrm{Z}_{w_i}^{(j)} = \mathrm{Attention}\left(\mathrm{Q}_{w_i}^{(j)}, \mathrm{K}_{w_i}^{(j)}, \mathrm{V}_{w_i}^{(j)} \right) = \mathrm{softmax}\left(\frac{\mathrm{Q}_{w_i}^{(j)} \left(\mathrm{K}_{w_i}^{(j)} \right)'}{\sqrt{C'}} \right) \mathrm{V}_{w_i}^{(j)} \tag{8}$$

where $\mathrm{Q}_{w_i}^{(j)}, \mathrm{K}_{w_i}^{(j)}, \mathrm{V}_{w_i}^{(j)}, \mathrm{Z}_{w_i}^{(j)} \in \mathbb{R}^{M^2 \times C'}$ and $C' = \frac{C}{h}$.

Finally, the features are concatenated to restore the original input shape.

The original window-based attention operation employs a fixed window size that is always horizontal and vertical. Using the coordinates (x_c, y_c), (x_l, y_l), and (x_r, y_r) to represent the center, upper left, and lower right pixels of the window, respectively, we have:

$$\begin{bmatrix} x_l \\ y_l \\ x_r \\ y_r \end{bmatrix} = \begin{bmatrix} x_c \\ y_c \\ x_c \\ y_c \end{bmatrix} + \begin{bmatrix} x_l^r \\ y_l^r \\ x_r^r \\ y_r^r \end{bmatrix} \tag{9}$$

where x_l^r, y_l^r, x_r^r, and y_r^r denote the distance between the coordinates of the corner points and the coordinates of the center point, respectively.

It is well known that remote sensing images often contain various target objects with arbitrary orientations and different scales. Therefore, a fixed and unchangeable window is not an optimal design. Unlike the original window-based attention operations, RVSA (as shown in Figure 6) does not rely on fixed-size window partitions at a fixed orientation. Instead, it produces windows with different positions, sizes, shapes, and angles by adjusting learnable shift, scale, and rotation parameters (O_w, S_w, and Θ_w, respectively). Specifically, distinct prediction layers can be employed for each window to estimate the shift, scale, and rotation parameters for *key* and *value* tokens, relying on the input features:

$$O_w^K, S_w^K, \Theta_w^K = \mathrm{Linear}_K(\mathrm{LeakyReLU}(\mathrm{GAP}(\mathrm{X}_w))) \tag{10}$$

$$O_w^V, S_w^V, \Theta_w^V = \mathrm{Linear}_V(\mathrm{LeakyReLU}(\mathrm{GAP}(\mathrm{X}_w))) \tag{11}$$

where GAP is the global average pooling operation.

Afterward, following the parameters mentioned earlier, the initial window undergoes transformation. The transformed coordinates of the corner points $\left(x'_{l/r}, y'_{l/r} \right)$ are then calculated using the following formulas:

$$\begin{bmatrix} x'_{l/r} \\ y'_{l/r} \end{bmatrix} = \begin{bmatrix} x^c \\ y^c \end{bmatrix} + \begin{bmatrix} o_x \\ o_y \end{bmatrix} + \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_{l/r}^r \cdot s_x \\ y_{l/r}^r \cdot s_y \end{bmatrix} \tag{12}$$

where o_x, o_y, s_x, s_y, and θ denote the shift, scale and rotation parameters. Namely,$O_w = \left\{ o_x, o_y \in \mathbb{R}^1 \right\}$, $S_w = \left\{ s_x, s_y \in \mathbb{R}^1 \right\}$, and $\Theta_w = \left\{ \theta \in \mathbb{R}^1 \right\}$.

Then, the *key* and *value* features are sampled from the transformed windows, which are then utilized to compute multi-head self-attention. As each head can generate windows with different positions, sizes, and shapes, RVSA is especially effective in extracting information from multiple target objects with diverse scales and orientations. In addition, RVSA not only reduces computational complexity linearly with respect to the image size, but also seamlessly integrates with the existing framework since the difference lies in the manner of attention calculation, which is parameter-free. Therefore, it is highly suitable for the image super-resolution of the RSA system.

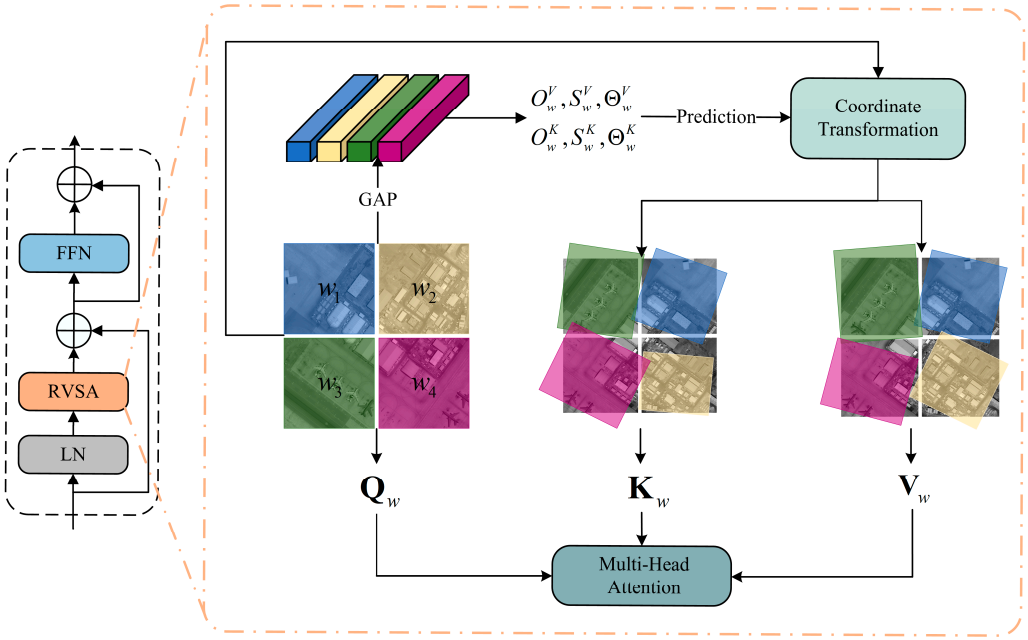


Figure 6. Illustration of rotated varied-size multi-head attention.

2.4. Decoder

The decoder is also composed of rotated varied-size window-based attention blocks. Its input consists of the extracted features of visible tokens and mask tokens, where each mask token is a shared, learned vector that indicates the presence of a missing patch to be predicted. To enhance its suitability for the image super-resolution task, our decoder matches the depth (number of Transformer blocks) and width (number of channels) of the encoder.

2.5. Implementation and Training Details

We employ the "base" version of the ViT as both the encoder and decoder. Specifically, the encoder and decoder have a depth of 12 and a width of 1024 d, respectively. The head number, patch size, embedding dimension, and multilayer perceptron ratio are set to 12, 16, 768, and 4, respectively. Following the strategy proposed by [36,41], we use the original Vision Transformer blocks at the 3rd, 6th, 9th, and 12th layers, while utilizing the rotated varied-size window-based attention blocks with a window size of 7 for the remaining layers.

During pre-training, we set the masking ratio to 75%, which is the same as the original MAE. To facilitate masking the image, for each image taken at different rotation angles of the rectangular pupil, we rotate it to the direction of the short side of the rectangle, which is horizontal or vertical, and perform center cropping. Following the guidelines provided in [37], the default settings for both pre-training and fine-tuning can be found in Table 2. To calculate the loss, we use the Charbonnier loss function $\mathcal{L} = \sqrt{\|R - G\|^2 + \varepsilon^2}$, with G representing the ground-truth high-quality image and R representing the SR result. The value of ε is set at 1×10^{-3}. All experiments were conducted on a workstation equipped

with an Intel i9-12900K CPU and an NVIDIA RTX 4080 GPU, each with a memory size of 16GB. The model underwent training for a total of 400 epochs, with each epoch requiring approximately 97 min.

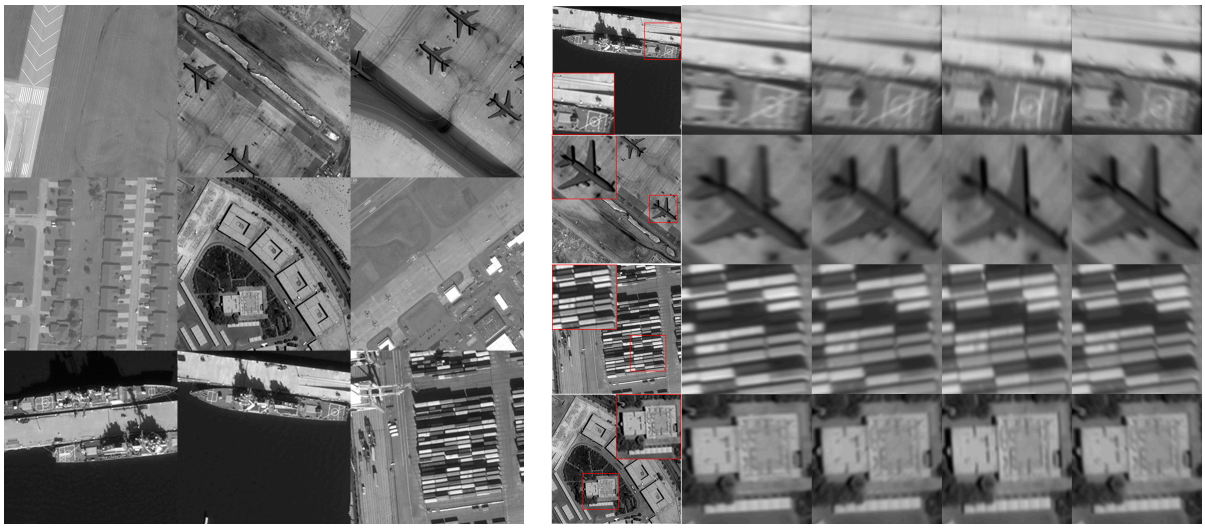
Table 2. Training settings.

Config	Pre-training	Fine-Tuning
Optimizer	AdamW [42]	
Base Learning Rate	1.5×10^{-4}	1×10^{-3}
Weight Decay	0.05	
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$	$\beta_1 = 0.9, \beta_2 = 0.999$
Batch Size	256	64
Learning Rate Schedule	Cosine Decay [43]	

3. Results

3.1. Experimental Setup

To validate the effectiveness of the proposed method, two types of experiments were conducted: digital simulation and semi-physical imaging simulation. The digital simulation experiment utilized high-resolution remote sensing images as inputs. These images were downsampled, and a full-link digital approach was employed to simulate the degradation of imaging quality in the RSA system [44]. This simulation generated degraded images, which were then used for constructing datasets. Specifically, we obtained a total of 210 original scene images from the WorldView-3 satellite data. For each scene, we conducted 24 sets of image simulations, encompassing six aspect ratios ranging from 3 to 8, and four primary mirror rotation angles: 0°, 45°, 90°, and 135°. As a result, the dataset consisted of a total of 5040 images. Of these, 90% were allocated for the training set, totaling 4536 images. Figure 7 illustrates some of the target scenes and their degraded images obtained through the digital simulation method. These images demonstrate that the resolution of the images acquired by the RSA system varies significantly in different directions, as evidenced by the markings on the decks of naval vessels, the edges of shipping containers, airplanes, and square buildings.



(**a**) (**b**)

Figure 7. (**a**) Target scenes; (**b**) degraded images with different rotation angles, from left to right, corresponding to rotation angles of 0°, 45°, 90°, and 135°.

In the semi-physical imaging simulation experiment, an imaging platform was utilized to simulate the RSA system's imaging process for imaging target scenes or resolution

targets [17]. The captured real images were used for testing purposes. Figure 8 displays the design scheme diagram, while Figure 9 displays the physical diagram.

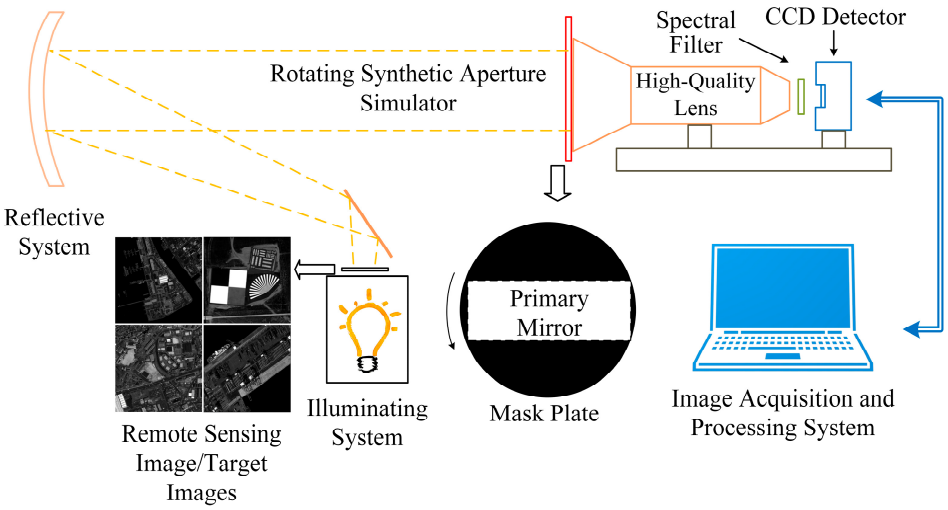


Figure 8. Design scheme of the imaging experiment platform.

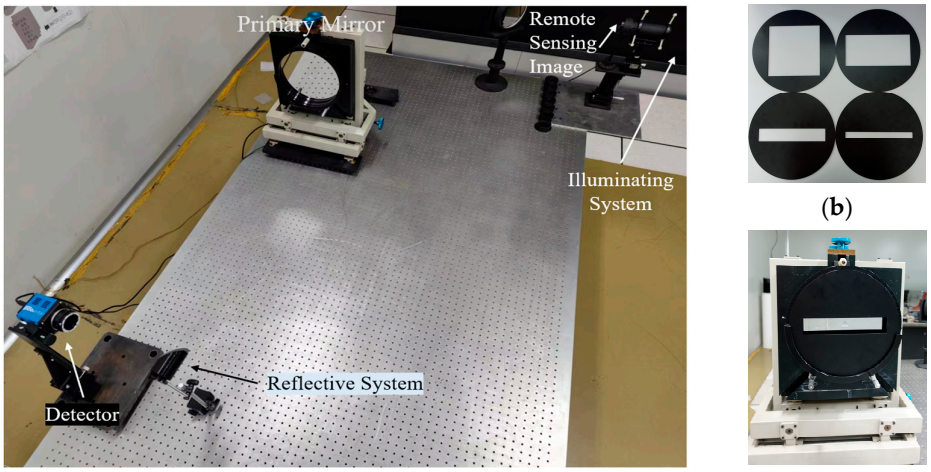


Figure 9. (**a**) Semi-physical imaging experiment platform; (**b**) rectangular pupil optical elements; (**c**) the primary mirror with a rectangular pupil optical element [44].

As there are currently no other SISR methods specifically designed for the RSA system, we compared our proposed method with general SISR methods. These methods include SRGAN, EDSR, SRMD, and RealESRGAN, which are representative explicit methods that use external training datasets, DualSR, which is an explicit method that uses internal statistics of images, and FSSR, which is a representative implicit method.

3.2. Experimental Results

3.2.1. Quantitative Results

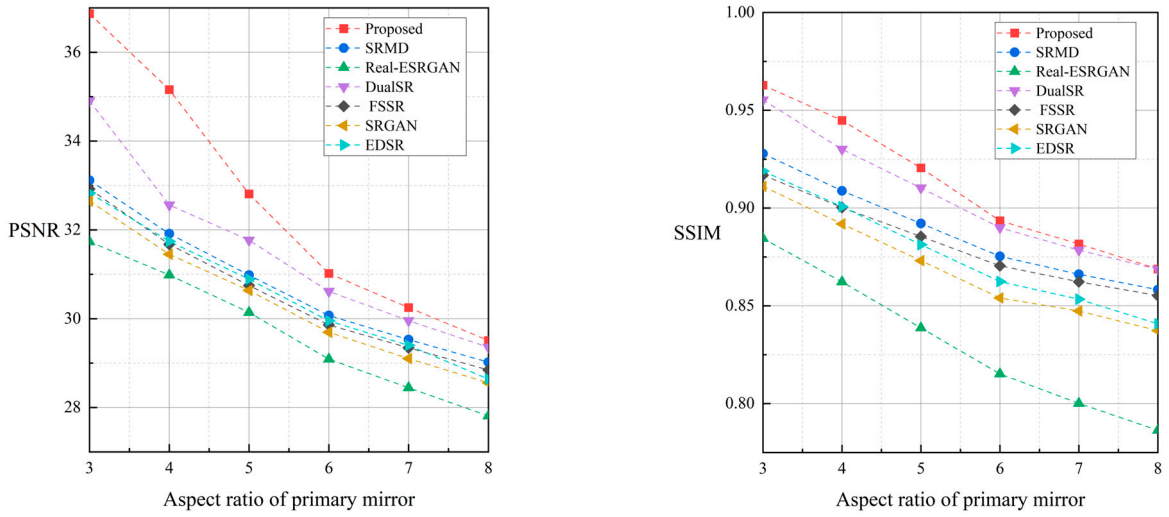
Table 3 presents a quantitative evaluation of the aforementioned methods, utilizing two quality metrics, namely, PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity) [45]. These metrics gauge the level of resemblance between two images. The table also includes the results obtained through bicubic interpolation for the purpose of comparison. It contains the super-resolution results for six different scenes, each with six different aspect ratios of the rectangular primary mirror. To provide a comprehensive overview, Table 4 and Figure 10 show the average results for all test images.

Table 3. Super-resolution results. The unit of PSNR is decibel (dB). The best result for each scene type and aspect ratio is highlighted in bold font.

Scene Type	Method	Aspect Ratio 3		Aspect Ratio 4		Aspect Ratio 5		Aspect Ratio 6		Aspect Ratio 7		Aspect Ratio 8	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Airports	Bicubic	28.63	0.8160	27.68	0.7992	26.87	0.7833	26.09	0.7677	25.73	0.7608	25.15	0.7505
	SRGAN	32.46	0.9022	31.41	0.8802	30.46	0.8573	29.37	0.8358	28.82	0.8237	28.07	0.8084
	EDSR	32.57	0.9195	31.52	0.8981	30.54	0.8782	29.42	0.8555	28.89	0.8455	28.02	0.8263
	SRMD	32.47	0.9228	31.31	0.9027	30.36	0.8841	29.46	0.8660	29.04	0.8580	28.37	0.8462
	Real-ESRGAN	32.68	0.9054	31.79	0.8829	30.77	0.8570	29.55	0.8273	28.89	0.8136	28.04	0.7952
	DualSR	34.22	0.9517	32.68	0.9294	31.22	0.9044	30.02	0.8812	**29.49**	0.8709	28.66	0.8553
	FSSR	32.28	0.9105	31.07	0.8926	30.13	0.8763	29.25	0.8603	28.85	0.8532	28.20	0.8427
	Proposed	**35.96**	**0.9584**	**34.68**	**0.9384**	**32.24**	**0.9142**	**30.40**	**0.8827**	29.31	**0.8745**	**28.80**	**0.8554**
Harbors	Bicubic	27.98	0.8351	26.77	0.8193	26.05	0.8073	25.41	0.7982	24.99	0.7914	24.49	0.7847
	SRGAN	31.37	0.9333	30.11	0.9156	29.27	0.8997	28.59	0.8891	28.06	0.8806	27.53	0.8732
	EDSR	31.40	0.9409	30.17	0.9233	29.26	0.9070	28.62	0.8958	28.16	0.8907	27.61	0.8816
	SRMD	31.77	0.9440	30.30	0.9253	29.44	0.9111	28.70	0.9007	28.20	0.8928	27.63	0.8851
	Real-ESRGAN	30.65	0.9187	29.48	0.9015	28.72	0.8840	28.02	0.8723	27.60	0.8640	27.05	0.8560
	DualSR	34.55	0.9668	31.89	0.9438	30.54	0.9252	29.45	0.9123	28.82	0.9028	28.12	0.8936
	FSSR	32.65	0.9433	30.88	0.9253	29.84	0.9109	29.02	0.9009	28.47	0.8932	27.85	0.8857
	Proposed	**36.38**	**0.9712**	**34.59**	**0.9544**	**31.79**	**0.9312**	**30.57**	**0.9132**	**29.39**	**0.9043**	**28.55**	**0.8941**
Residential areas	Bicubic	28.42	0.8031	27.69	0.7853	26.55	0.7636	25.81	0.7436	25.37	0.7326	24.84	0.7219
	SRGAN	31.59	0.8844	30.59	0.8669	29.81	0.8394	29.02	0.8181	28.31	0.8168	27.73	0.8055
	EDSR	31.92	0.8948	31.04	0.8784	30.14	0.8472	29.30	0.8270	28.74	0.8213	28.07	0.7993
	SRMD	32.23	0.9095	31.34	0.8880	30.01	0.8626	29.14	0.8393	28.63	0.8267	28.01	0.8143
	Real-ESRGAN	32.15	0.8839	31.32	0.8547	29.85	0.8113	28.73	0.7794	28.04	0.7592	27.22	0.7382
	DualSR	33.92	0.9455	32.73	0.9216	30.85	0.8892	29.67	0.8583	29.03	0.8428	28.32	0.8279
	FSSR	32.10	0.9006	31.14	0.8807	29.88	0.8579	29.00	0.8362	28.49	0.8243	27.89	0.8128
	Proposed	**36.36**	**0.9557**	**34.85**	**0.9408**	**31.83**	**0.9026**	**30.35**	**0.8679**	**29.52**	**0.8499**	**28.48**	**0.8316**
Yards	Bicubic	27.19	0.8143	26.31	0.7987	25.50	0.7833	24.67	0.7677	24.44	0.7629	24.03	0.7563
	SRGAN	30.67	0.9140	29.75	0.8937	28.96	0.8715	27.86	0.8415	27.60	0.8453	26.89	0.8268
	EDSR	30.73	0.9133	29.80	0.8978	29.01	0.8786	27.90	0.8521	27.65	0.8467	27.04	0.8342
	SRMD	30.86	0.9212	29.80	0.9025	28.84	0.8847	27.87	0.8666	27.60	0.8612	27.12	0.8535
	Real-ESRGAN	30.22	0.8902	29.44	0.8692	28.81	0.8494	27.58	0.8193	27.33	0.8130	26.58	0.7963
	DualSR	32.68	0.9467	27.77	0.9124	29.67	0.9006	28.54	0.8808	28.14	0.8736	27.58	0.8641
	FSSR	31.37	0.9145	30.11	0.8969	29.06	0.8806	28.01	0.8642	27.70	0.8590	27.20	0.8519
	Proposed	**35.63**	**0.9621**	**33.47**	**0.9395**	**30.87**	**0.9155**	**29.18**	**0.8836**	**28.46**	**0.8756**	**28.14**	**0.8684**
Farmland	Bicubic	32.60	0.8556	31.35	0.8472	30.74	0.8426	29.93	0.8367	29.19	0.8319	28.71	0.8285
	SRGAN	36.33	0.9586	34.90	0.9505	34.25	0.9455	33.35	0.9389	32.58	0.9344	32.06	0.9309
	EDSR	36.70	0.9620	35.22	0.9534	34.49	0.9472	33.54	0.9406	32.82	0.9360	32.25	0.9326
	SRMD	37.05	0.9651	35.52	0.9553	34.81	0.9500	33.85	0.9432	32.98	0.9376	32.42	0.9338
	Real-ESRGAN	31.73	0.8473	31.71	0.8376	31.28	0.8308	30.21	0.8221	29.27	0.8110	28.75	0.8076
	DualSR	37.43	0.9717	35.78	0.9610	34.91	0.9533	**33.88**	0.9466	33.00	0.9410	**32.46**	**0.9373**
	FSSR	35.62	0.9522	34.29	0.9458	33.69	0.9411	32.85	0.9346	32.18	0.9311	31.70	0.9280
	Proposed	**38.85**	**0.9766**	**37.05**	**0.9645**	**35.66**	**0.9590**	33.25	**0.9483**	**33.34**	**0.9430**	31.92	0.9292
Forests	Bicubic	30.30	0.7985	29.38	0.7772	28.70	0.7612	27.85	0.7409	27.24	0.7271	27.10	0.7237
	SRGAN	33.47	0.8739	31.95	0.8446	31.07	0.8252	29.98	0.8008	29.24	0.7835	29.05	0.7788
	EDSR	33.65	0.8828	32.69	0.8533	31.92	0.8299	30.95	0.8037	30.14	0.7803	29.90	0.7706
	SRMD	34.34	0.9045	33.22	0.8791	32.43	0.8601	31.43	0.8364	30.73	0.8204	30.57	0.8164
	Real-ESRGAN	32.96	0.8610	32.15	0.8276	31.41	0.7998	30.47	0.7710	29.55	0.7402	29.22	0.7248
	DualSR	36.65	0.9493	34.52	0.9120	33.44	0.8892	32.12	0.8607	31.25	0.8399	31.00	0.8340
	FSSR	33.50	0.8797	32.56	0.8604	31.90	0.8459	31.01	0.8265	30.37	0.8129	30.22	0.8100
	Proposed	**38.04**	**0.9530**	**36.33**	**0.9313**	**34.48**	**0.9005**	**32.39**	**0.8650**	**31.50**	**0.8427**	**31.19**	**0.8341**

Table 4. Average super-resolution results. The unit of PSNR is decibel (dB). The best result is highlighted in bold font.

Scene Type	Method	Aspect Ratio 3		Aspect Ratio 4		Aspect Ratio 5		Aspect Ratio 6		Aspect Ratio 7		Aspect Ratio 8	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Average	Bicubic	29.19	0.8204	28.20	0.8045	27.40	0.7902	26.63	0.7758	26.16	0.7678	25.72	0.7609
	SRGAN	32.65	0.9111	31.45	0.8919	30.64	0.8731	29.70	0.8540	29.10	0.8474	28.55	0.8373
	EDSR	32.83	0.9189	31.74	0.9007	30.89	0.8813	29.95	0.8624	29.40	0.8534	28.64	0.8408
	SRMD	33.12	0.9278	31.91	0.9088	30.98	0.8921	30.07	0.8754	29.53	0.8661	29.02	0.8582
	Real-ESRGAN	31.73	0.8844	30.98	0.8622	30.14	0.8387	29.09	0.8152	28.45	0.8002	27.81	0.7864
	DualSR	34.91	0.9553	32.56	0.9300	31.77	0.9103	30.61	0.8900	29.95	0.8785	29.36	0.8687
	FSSR	32.92	0.9168	31.68	0.9003	30.75	0.8854	29.86	0.8705	29.34	0.8623	28.85	0.8552
	Proposed	**36.87**	**0.9628**	**35.16**	**0.9448**	**32.81**	**0.9205**	**31.02**	**0.8935**	**30.25**	**0.8817**	**29.51**	**0.8688**



(**a**) (**b**)

Figure 10. Average super-resolution results. (**a**) PSNR; (**b**) SSIM.

3.2.2. Qualitative Results

In addition to the quantitative evaluations provided by the full-reference metrics mentioned earlier, we present visual results as qualitative evaluations for the scenes depicted in Figure 7. Figures 11–14 illustrate the super-resolution outcomes for SRGAN, EDSR, SRMD, real-ESRGAN, DualSR, FSSR, and the proposed method. Additionally, Figure 15 displays semi-physical imaging experimental images with a primary mirror aspect ratio of 3 and a rotation angle of 90. Specifically, the local enlargement image is shown in Figure 15a, while the SR results generated by SRGAN, EDSR, SRMD, real-ESRGAN, DualSR, FSSR, and the proposed method are displayed in Figure 15b–h, respectively.

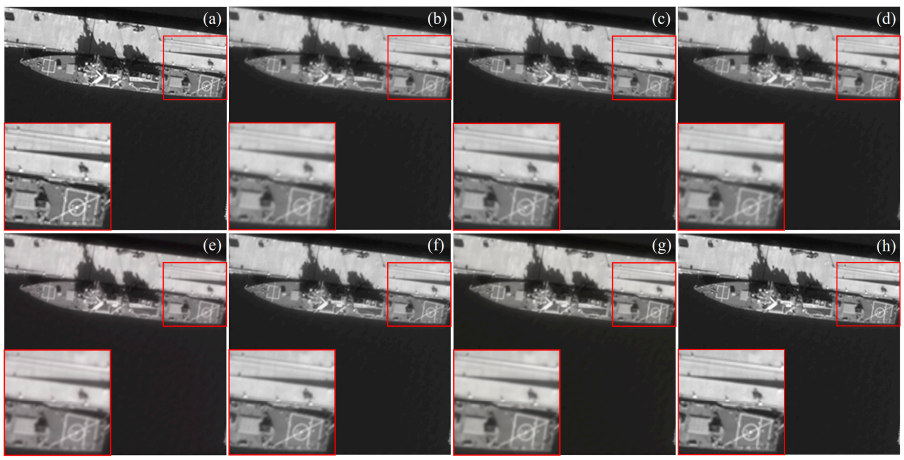


Figure 11. HR and SR results of the test image harbor with a rotation angle of 45° and an aspect ratio of 3. (**a**) HR; (**b**) SRGAN; (**c**) EDSR; (**d**) SRMD; (**e**) real-ESRGAN; (**f**) DualSR; (**g**) FSSR; and (**h**) proposed method.

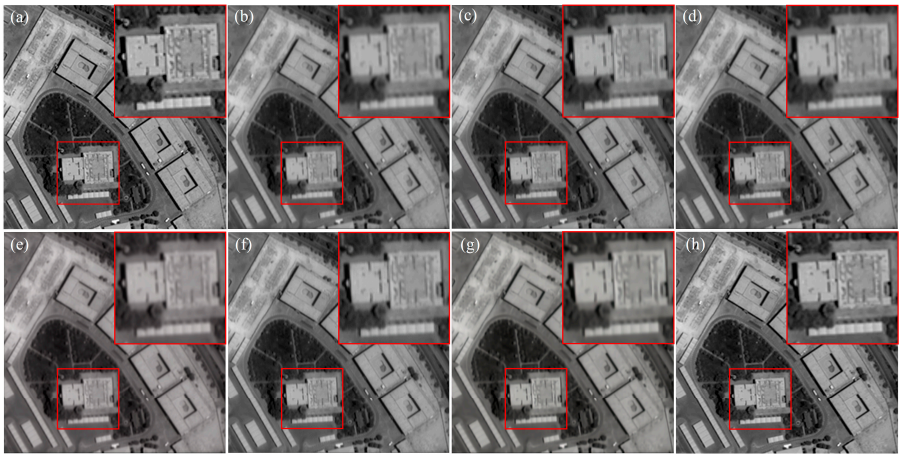


Figure 12. HR and SR results of the test image residential area with a rotation angle of 90° and an aspect ratio of 4. (**a**) HR; (**b**) SRGAN; (**c**) EDSR; (**d**) SRMD; (**e**) real-ESRGAN; (**f**) DualSR; (**g**) FSSR; and (**h**) proposed method.

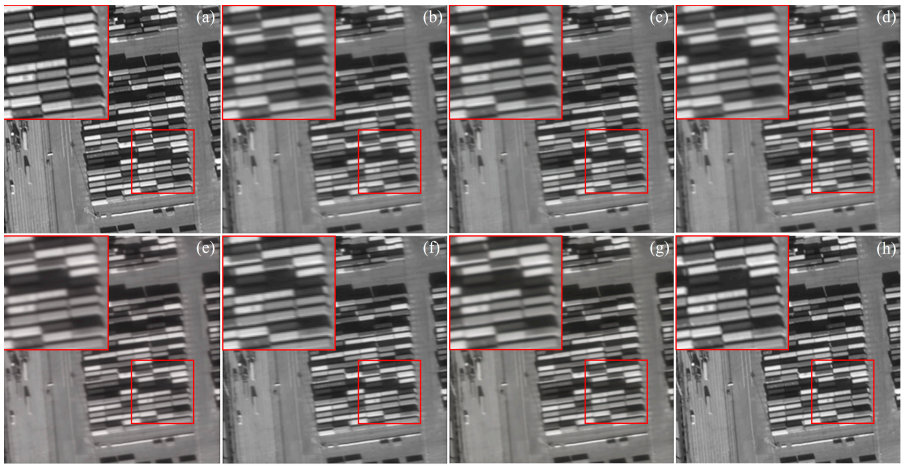


Figure 13. HR and SR results of the test image yard with a rotation angle of 0° and an aspect ratio of 5. (**a**) HR; (**b**) SRGAN; (**c**) EDSR; (**d**) SRMD; (**e**) real-ESRGAN; (**f**) DualSR; (**g**) FSSR; and (**h**) proposed method.

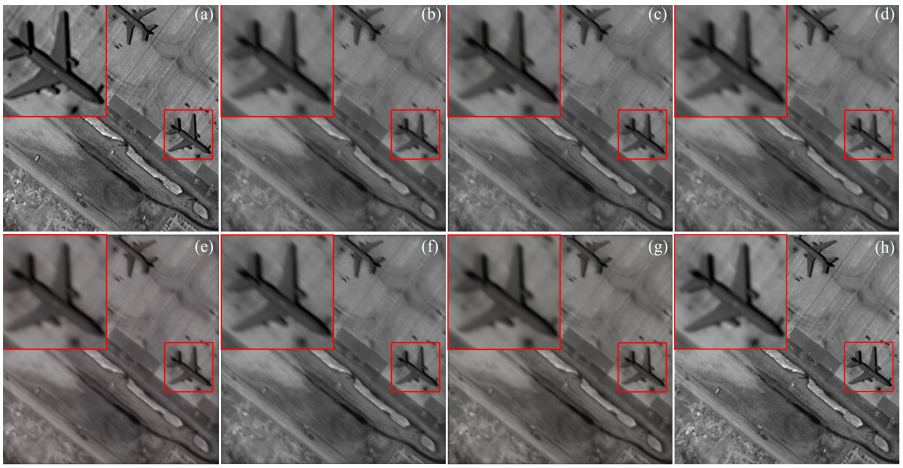


Figure 14. HR and SR results of the test image airport with a rotation angle of 135° and an aspect ratio of 6. (**a**) HR; (**b**) SRGAN; (**c**) EDSR; (**d**) SRMD; (**e**) real-ESRGAN; (**f**) DualSR; (**g**) FSSR; and (**h**) proposed method.

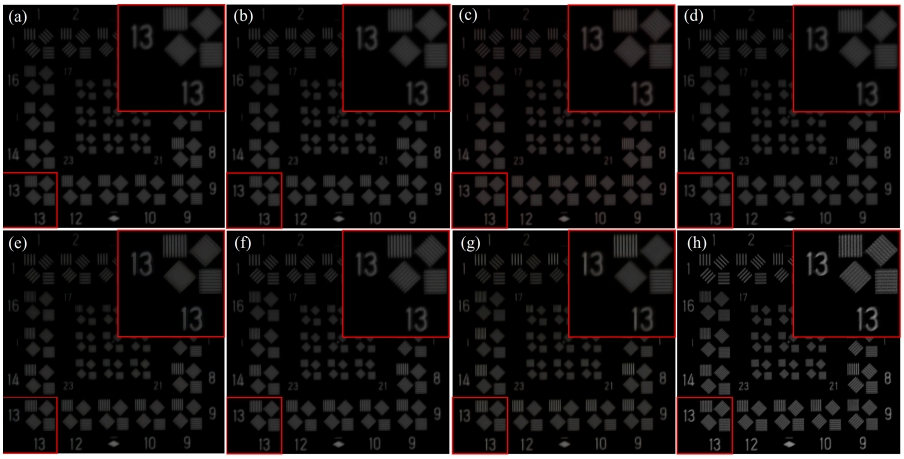


Figure 15. LR and SR results of the semi-physical experimental image with a rotation angle of 0° and an aspect ratio of 3. (**a**) LR; (**b**) SRGAN; (**c**) EDSR; (**d**) SRMD; (**e**) real-ESRGAN; (**f**) DualSR; (**g**) FSSR; and (**h**) proposed method.

3.2.3. Ablation Study

We compared two training methods: the "one-stage" method, which involves directly training on unmasked degraded images for 400 epochs, and the "two-stage" method, which involves first pre-training on images masked with strip-like masks along the direction of the rectangle's short side for 360 epochs, followed by fine-tuning on the unmasked images for 40 epochs. Table 5 presents the results of the ablation study, which includes the average results of all target scenes with six different aspect ratios ranging from 3 to 8. The table also includes a comparison with the results obtained from randomly masked images for pre-training.

Table 5. Ablation study on the three training methods. The unit of PSNR is decibel (dB).

Training Method	One-Stage	Two-Stage-Random Mask	Two-Stage-Strip Mask
PSNR	32.07	31.91	32.61
SSIM	0.9038	0.9024	0.9120

4. Discussion

The main objective of image super-resolution for the RSA system is to enhance the resolution in the direction of the shorter edge of the rectangular pupil. In contrast, the resolution of LR and HR in the longer edge direction is essentially the same. As such, experimental results demonstrate that a straightforward ViT model can produce outstanding results without requiring overly complex structures. Table 3 lists 36 sets of digital simulation test images of six scenes and six aspect ratios. The proposed method achieves the best performance in 33 sets of test images according to the PSNR metric and in 35 sets according to the SSIM metric. While DualSR outperforms the proposed method in some scenes, the proposed method yields significantly superior average results. For different aspect ratios of the primary mirror, the proposed method performs exceptionally well when the aspect ratio is 3, with a 26.31% improvement in PSNR and a 17.36% improvement in SSIM over bicubic interpolation. Similarly, when the aspect ratio of the primary mirror is 4 or 5, our proposed method achieves the best results and significantly outperforms other methods. Even when the aspect ratio of the primary mirror is 6, our super-resolution results still exhibit a PSNR of over 31dB, owing to the proposed method's consideration of the imaging characteristics of the RSA system. However, the performance of the proposed method decreases when the aspect ratio is large (greater than or equal to 7). We attribute this issue to the high degree of blurriness that is prevalent along the shorter side of the rectangular pupil. Despite applying strip-shaped masking to the images, numerous blurred

edges remain challenging to conceal. Hence, these blurred pixels may still impede the model's super-resolution reconstruction.

From the visual results in Figures 11–15, it is evident that each SR method tends to exhibit specific visual characteristics in the SR output, which can be classified into two categories. One category, exemplified by SRGAN and Real-ESRGAN, tends to generate smoother outputs with clearer visual effects, making them more robust against noise. However, these methods exhibit subpar performance on objective evaluation metrics. The other methods tend to produce sharper SR outputs. Furthermore, the visual results show that when the aspect ratio of the primary mirror is relatively large, the image quality along the shorter side is significantly reduced. Although some details can be restored, the SR results may still fall short of meeting the resolution requirements of interpretation applications. Compared to other methods, our proposed method exhibits reduced susceptibility to the adverse effects of non-uniform resolution and demonstrates a superior ability to reconstruct directions with low resolution in the image. This is particularly evident from the outcomes of our semi-physical imaging simulation experiments shown in Figure 15. In addition, the fact that the semi-physical experimental test image and the training image were obtained from different sensors also serves to demonstrate the robust generalization capability of our method.

The results of ablation experiments demonstrate the advantages of our "two-stage" training method. As shown in Table 5, the performance of the model pre-trained on randomly masked images is similar to that of the "one-stage" training method. However, our "two-stage" training method, which utilizes a strip-wise mask sampling strategy, improves the SSIM and PSNR by 0.91% and 1.68%, respectively, compared to the "one-stage" method. This outcome is expected because remote sensing images possess a significant amount of spatial redundancy, which means that even if some pixels are masked, deep neural networks are capable of extracting enough information from the images to infer complex and holistic reconstructions. Furthermore, as explained in Section 2.3.1, masking along the shorter side of the rectangle can help the model avoid interference from blurred pixels, resulting in reconstructed images that are sharper and clearer along the shorter edge.

5. Conclusions

In this paper, we conduct an analysis of the imaging characteristics of the RSA system and put forth a corresponding SISR method. Our proposed method employs an end-to-end image super-resolution network that is based on the rotated varied-size window-based attention mechanism. By utilizing window-based self-attention, this mechanism generates windows with varying locations, sizes, shapes, and angles. Such an approach proves advantageous in effectively processing objects with diverse orientations and scales in remote sensing images. To effectively handle the special asymmetric degradation characteristic of the RSA system, we employ a mask strategy using strip-wise masks along the short side of the rectangular primary mirror. On this basis, we adopt a two-stage training method that involves pre-training the model on masked images, followed by fine-tuning using unmasked images. This approach not only mitigates interference caused by the non-circular symmetry PSF but also enhances the network's ability to make more effective use of the high-resolution information inherent in the remote sensing images themselves. Consequently, our network excels in reconstructing detailed and clear edges and textures in the direction of the shorter edge of the pupil. Extensive experiments are conducted, which include six aspect ratios of the primary mirror and six different SR methods, to demonstrate the superior performance of our proposed method. Specifically, our method outperforms other methods in objective evaluation for primary mirrors with aspect ratios ranging from 3 to 8, especially in terms of the PSNR metric. Furthermore, our method effectively addresses the issue of uneven resolution in SR results, showcasing its superiority in image interpretation applications. Through this research, we offer valuable guidance for the practical implementation of the RSA imaging technology, while also providing significant references for its future advancements.

Author Contributions: Conceptualization, Y.S. and S.J.; methodology, Y.S. and S.J.; software, Y.S. and T.S.; validation, X.Z. and S.J.; formal analysis, S.J. and T.S.; investigation, J.S. and J.Y.; resources, X.Z.; data curation, T.S. and J.Y.; writing—original draft preparation, Y.S. and S.W.; writing—review and editing, Y.S. and S.W.; visualization, Y.S. and J.S.; supervision, X.Z. and W.Z.; project administration, X.Z. and W.Z.; funding acquisition, X.Z. and S.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (NSFC) (62305086, 62101160, and 61975043) and the China Postdoctoral Science Foundation (2023M740901).

Data Availability Statement: The datasets used or analyzed during the current study are available from the corresponding author upon reasonable request.

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

References

1. Yang, X.; Li, F.; Xin, L.; Lu, X.; Lu, M.; Zhang, N. An improved mapping with super-resolved multispectral images for geostationary satellites. *Remote Sens.* **2020**, *12*, 466. [CrossRef]
2. Yao, L.; Liu, Y.; He, Y. A Novel Ship-Tracking Method for GF-4 Satellite Sequential Images. *Sensors* **2018**, *18*, 2007. [CrossRef]
3. Kulkarni, S.C.; Rege, P.P. Pixel Level Fusion Techniques for SAR and Optical Images: A Review. *Inf. Fusion* **2020**, *59*, 13–29. [CrossRef]
4. Yu, W.; You, H.; Lv, P.; Hu, Y.; Han, B. A Moving Ship Detection and Tracking Method Based on Optical Remote Sensing Images from the Geostationary Satellite. *Sensors* **2021**, *21*, 7547. [CrossRef] [PubMed]
5. Zhang, Y.; Ye, M.; Zhu, G.; Liu, Y.; Guo, P.; Yan, J. FFCA-YOLO for Small Object Detection in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5611215. [CrossRef]
6. Zhao, J. Higher Temporal Evapotranspiration Estimation with Improved SEBS Model from Geostationary Meteorological Satellite Data. *Sci. Rep.* **2019**, *9*, 14981. [CrossRef] [PubMed]
7. Wang, Y.; Zhang, C.; Guo, L.; Xu, S.; Ju, G. Decoupled Object-Independent Image Features for Fine Phasing of Segmented Mirrors Using Deep Learning. *Remote Sens.* **2022**, *14*, 4681. [CrossRef]
8. Jikuya, I.; Uchida, D.; Kino, M.; Kurita, M.; Yamada, K. Development status of the segmented mirror control system in Seimei Telescope. In *Advances in Optical and Mechanical Technologies for Telescopes and Instrumentation IV*; SPIE: Montréal, QC, Canada, 2020; Volume 11451, pp. 965–973.
9. Atcheson, P.; Domber, J.; Whiteaker, K.; Britten, J.A.; Dixit, S.N.; Farmer, B. *MOIRE: Ground Demonstration of a Large Aperture Diffractive Transmissive Telescope*; Oschmann, J.M., Clampin, M., Fazio, G.G., MacEwen, H.A., Eds.; SPIE: Montréal, QC, Canada, 2014; p. 91431W.
10. Liu, D.; Wang, L.; Yang, W.; Wu, S.; Fan, B.; Wu, F. Stray Light Characteristics of the Diffractive Telescope System. *Opt. Eng.* **2018**, *57*, 1. [CrossRef]
11. Peng, Y.; Fu, Q.; Amata, H.; Su, S.; Heide, F.; Heidrich, W. Computational Imaging Using Lightweight Diffractive-Refractive Optics. *Opt. Express* **2015**, *23*, 31393. [CrossRef]
12. Tang, J.; Wang, K.; Ren, Z.; Zhang, W.; Wu, X.; Di, J.; Liu, G.; Zhao, J. RestoreNet: A deep learning framework for image restoration in optical synthetic aperture imaging system. *Opt. Lasers Eng.* **2021**, *139*, 106463. [CrossRef]
13. Rai, M.R.; Rosen, J. Optical incoherent synthetic aperture imaging by superposition of phase-shifted optical transfer functions. *Opt. Lett.* **2021**, *46*, 1712–1715. [CrossRef] [PubMed]
14. Wu, J.; Yang, F.; Cao, L. Resolution enhancement of long-range imaging with sparse apertures. *Opt. Lasers Eng.* **2022**, *155*, 107068. [CrossRef]
15. Sun, Y.; Zhi, X.; Jiang, S.; Fan, G.; Yan, X.; Zhang, W. Image Fusion for the Novelty Rotating Synthetic Aperture System Based on Vision Transformer. *Inf. Fusion* **2024**, *104*, 102163. [CrossRef]
16. Zhi, X.; Jiang, S.; Zhang, L.; Wang, D.; Hu, J.; Gong, J. Imaging mechanism and degradation characteristic analysis of novel rotating synthetic aperture system. *Opt. Lasers Eng.* **2021**, *139*, 106500. [CrossRef]
17. Sun, Y.; Zhi, X.; Zhang, L.; Jiang, S.; Shi, T.; Wang, N.; Gong, J. Characterization and Experimental Verification of the Rotating Synthetic Aperture Optical Imaging System. *Sci. Rep.* **2023**, *13*, 17015. [CrossRef]
18. Zhi, X.; Jiang, S.; Zhang, L.; Hu, J.; Yu, L.; Song, X.; Gong, J. Multi-frame image restoration method for novel rotating synthetic aperture imaging system. *Results Phys.* **2021**, *23*, 103991. [CrossRef]
19. Gendy, G.; He, G.; Sabor, N. Lightweight Image Super-Resolution Based on Deep Learning: State-of-the-Art and Future Directions. *Inf. Fusion* **2023**, *94*, 284–310. [CrossRef]
20. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Wang, Y.; Zhang, L. From Degrade to Upgrade: Learning a Self-Supervised Degradation Guided Adaptive Network for Blind Remote Sensing Image Super-Resolution. *Inf. Fusion* **2023**, *96*, 297–311. [CrossRef]
21. Wei, S.; Cheng, H.; Xue, B.; Shao, X.; Xi, T. Low-Cost and Simple Optical System Based on Wavefront Coding and Deep Learning. *Appl. Opt.* **2023**, *62*, 6171. [CrossRef]
22. Freeman, W.T.; Jones, T.R.; Pasztor, E.C. Example-based super-resolution. *IEEE Comput. Graph. Appl.* **2002**, *22*, 56–65. [CrossRef]

23. Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; Dong, C. Blind image super-resolution: A survey and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5461–5480. [CrossRef] [PubMed]
24. Chen, H. Real-World Single Image Super-Resolution: A Brief Review. *Inf. Fusion* **2022**, *79*, 124–145. [CrossRef]
25. Lepcha, D.C.; Goyal, B.; Dogra, A.; Goyal, V. Image Super-Resolution: A Comprehensive Review, Recent Trends, Challenges and Applications. *Inf. Fusion* **2023**, *91*, 230–260. [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 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
27. 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.
28. 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.
29. 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, Montreal, BC, Canada, 11–17 October 2021; pp. 1905–1914.
30. Bell-Kligler, S.; Shocher, A.; Irani, M. Blind super-resolution kernel estimation using an internal-gan. *arXiv* **2019**, arXiv:1909.06581.
31. Emad, M.; Peemen, M.; Corporaal, H. Dualsr: Zero-shot dual learning for real-world super-resolution. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1630–1639.
32. Kim, J.; Jung, C.; Kim, C. Dual back-projection-based internal learning for blind super-resolution. *IEEE Signal Process. Lett.* **2020**, *27*, 1190–1194. [CrossRef]
33. Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; Lin, L. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 701–710.
34. Fritsche, M.; Gu, S.; Timofte, R. Frequency separation for real-world super-resolution. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 3599–3608.
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
36. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer towards remote sensing foundation model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 5607315. [CrossRef]
37. He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; Girshick, R. Masked autoencoders are scalable vision learners. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 16000–16009.
38. 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 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1874–1883.
39. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
40. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
41. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring plain vision transformer backbones for object detection. In Proceedings of the Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part IX. Springer Nature: Cham, Switzerland, 2022; pp. 280–296.
42. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
43. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
44. Sun, Y.; Zhi, X.; Jiang, S.; Gong, J.; Shi, T.; Wang, N. Imaging Simulation Method for Novel Rotating Synthetic Aperture System Based on Conditional Convolutional Neural Network. *Remote Sens.* **2023**, *15*, 688. [CrossRef]
45. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]